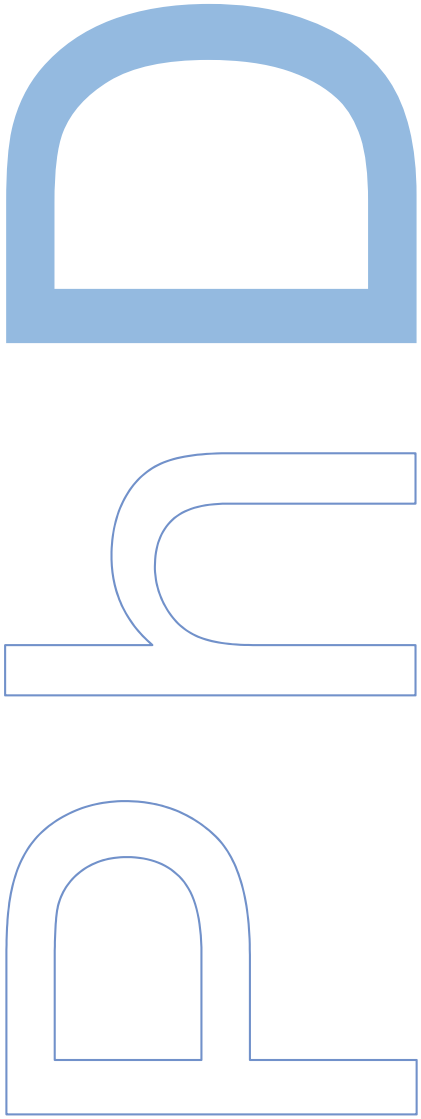
The Role of Non-coding DNA Structural Information in Phylogeny, Evolution and Disease

João Miguel Sotto Maior Faria Carneiro
Biologia
Departamento de Biologia da Faculdade de Ciências do Porto
2013

Orientador
António Manuel Amorim dos Santos, Professor Catedrático,
Faculdade de Ciências da Universidade do Porto

Coorientador
Maria João Ribeiro Nunes Ramos, Professora Catedrática,
Faculdade de Ciências da Universidade do Porto



ACKNOWLEDGEMENTS/AGRADECIMENTOS

Agradeço ao meu orientador António Amorim pela disponibilidade que sempre demonstrou para debater as ideias relacionadas com esta tese. Sobretudo agradeço a confiança que sempre depositou em mim. À minha co-orientadora, Maria João Ramos, um agradecimento pelo apoio e disponibilidade que demonstrou para discutir temas cruciais para a tese.

Agradeço a todas as pessoas que desenvolvem o seu trabalho no IPATIMUP e colaboraram de alguma forma comigo, pois parte da tese foi desenvolvida nas instalações e com os recursos deste instituto. Ao grupo de bioquímica teórica e computacional da Faculdade de Ciências da Universidade do Porto gratifico o apoio proporcionado em relação aos meios computacionais e apoio em dúvidas informáticas. Um agradecimento especial à Irina Moreira por todo o apoio que me deu.

A todo o grupo de genética populacional que desenvolve o seu trabalho no IPATIMUP um agradecimento pelo apoio e disponibilidade para tirar dúvidas que foram surgindo. Um agradecimento especial ao Filipe Pereira; à Luísa Azevedo, à Raquel Silva, ao Rune Matthiesen e ao Ricardo Araújo pelo apoio constante.

Agradeço ainda, a toda a minha família, especialmente ao meu pai e à minha mãe. Para a minha esposa, Selma, que esteve sempre a meu lado, agradeço a confiança que deposita em mim. Aos seus pais, Joaquim e Celeste também um agradecimento especial. A toda a minha família e amigos, obrigado pelo apoio.

Tese submetida à Faculdade de Ciências da Universidade do Porto para obtenção do grau de Doutor em Biologia

Thesis presented for the Doctor Degree in Biology in the Faculty of Sciences, University of Porto

1. Summary

Non-coding deoxyribonucleic acid (DNA) regions represent approximately 98% of the human genome and a relevant part of mitochondrial DNA (mtDNA). There is a clear contrast between coding and non-coding DNA regions considering the levels of genetic diversity, genomic architecture and distribution of regulatory elements. By using recently developed methodologies to analyse DNA, the unique features of coding regions and non-coding regions were accessed. For this purpose, four genetic models were used in this thesis: a) metallothioneins (MT), where specific mutational events converted a transcribed coding region into a non-coding region; b) Nicotinamidases (PNCs) and Nicotinamide phosphoribosyltransferases (NAMPTs) genes which presented critical structural hotspots related with the functionality of the respective proteins, and might have implications in the maintenance of expressed coding regions; c) non-coding mtDNA regions, and d) non-coding short tandem repeats (STRs).

The contrasts between coding (protein genes) and non-coding region (pseudogenes) were focused using a phylogenetic analysis associated to duplicated genes (model a). Mammalian evolution history of post-duplication events was herein explored by the study of MT family members where different mutational events can determine the way to a new function or to pseudogenisation.

Analysis of NAMPTs and PNCs (model b) homologous genes in different species was used to establish the relationships between mutations occurred during evolution and their consequences in metabolic pathways and pathologic conditions (e.g., cancer). The critical residues at active site and at the interaction with the substrate of invertebrate NAMPTs, nicotinamide, were maintained, considering both protein-docking analysis and expression. Nevertheless, additional hydrogen bonds and hydrophobic contacts were found in PNCs, what can be explained from complementary amino acid changes as a result of epistatic (compensatory) interactions. Structural conservation validated by expression experimental data was used to ascertain the current functional status and the evolutionary time depth of transcriptional loss of both NAMPT and PNC proteins in different species. This was useful to understand the molecular behaviour of specific chemical bonds (e.g., H-bonds) in proteins, which were also analysed in the DNA non-B conformations (model c and d) localized in the non-coding regions, even though they represent different types of molecules. By this way the computational molecular systems knowledge applied to proteins can be used to build the models for the DNA structures found in non-coding regions.

The study of conformational structural changes in non-B DNA conformations is very important since, as in proteins, they can adopt different structures related with specific properties. Furthermore, the genome architecture (coding versus non-coding) led us to the

analysis of the specificities of non-B conformations formation in mtDNA complete genome and their implications in biological processes (model c). Non-coding regions were playing a critical role in the process of generating different mtDNA deleted molecules associated with disease.

Ultimately, a new methodology for detection of secondary and tertiary DNA structures in non-coding regions was developed (model d). Available data for Y-chromosome short tandem repeats (STRs) was investigated by using software for structures prediction and new algorithms to identify non-B DNA conformations. Evaluation of these structures was attempted using molecular dynamics simulations and molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) calculations. Single-stranded and UNAFold predicted DNA conformations were analysed using chemical computational methodologies. Molecular structural features present in nuclear DNA (STRs) were inferred and correlated with different biological processes and diseases. Our analysis predicted hairpins that can arise in single-stranded STRs. The occurrence of these non-B DNA conformations in non-coding regions might influence/regulate processes of transcription occurring in protein-coding regions, and processes that depend of specific folding potential as DNA replication.

There was a clear contrast between protein-coding (model a and b) and non-coding genomes (model c and d). The possibility of these two different regions to generate or form three-dimensional structural molecules was accessed. The relevant DNA non-B conformations can adopt different conformations, as in proteins molecular systems, and was demonstrated in this thesis. In non-coding regions, the formation of DNA non-B conformations has implications in evolution, deletions, replication and disease (models c and d).

Sumário

As regiões não-codificantes de ácido desoxirribonucleico (ADN) representam cerca de 98% do genoma humano e de uma parte relevante do ADN mitocondrial (ADNmt). Há um contraste claro entre as regiões codificantes e regiões não codificantes do ADN considerando os níveis de diversidade genética, a arquitetura genómica e de distribuição de elementos de regulação. Utilizando metodologias recentemente desenvolvidas para a análise de ADN, as características únicas de regiões codificantes e não-codificantes foram determinadas. Para este efeito, quatro modelos genéticos foram utilizados neste trabalho: a) metalotioneínas (MT), onde padrões específicos de mutação podem converter uma região transcrita em uma região não codificante, b) os genes codificando as enzimas, nicotinamidase (PNCs) e nicotinamida phosphoribosyltransferases (NAMPTs), que apresentam 'hotspots' estruturais críticos relacionadas com a funcionalidade das proteínas respetivas, o que tem implicações na manutenção das regiões codificantes expressas; c) as regiões não codificantes do ADNmt, e d) as regiões não codificantes repetitivas em microssatélites.

Usando o modelo A, os contrastes entre as regiões codificante (genes) e não codificante (pseudogenes) foram analisados utilizando uma análise filogenética associada a genes duplicados. A evolução dos eventos pós duplicação dos genes MT nos mamíferos foi explorada pelo estudo de diferentes eventos mutacionais que podem determinar se um gene é ou não funcional.

A análise dos genes homólogos NAMPTs e PNCs (modelo b) em diferentes espécies foi usada para estabelecer as relações entre os resíduos resultantes de mutações durante a evolução e as suas consequências para as vias metabólicas e condições patológicas (por exemplo, cancro). Os resíduos críticos do centro ativo e interações de NAMPTs com o substrato, a nicotinamida, foram mantidos, considerando tanto a análise de 'docking' como a expressão das proteínas. No entanto, ligações de hidrogénio e contactos hidrofóbicos adicionais foram encontrados em PNCs, o que pode ser explicado a partir de alterações de aminoácidos complementares, como resultado de interações epistáticas. A conservação estrutural validada pelos dados experimentais de expressão foi usada para avaliar o estado funcional e a profundidade do tempo evolutivo de perda de transcrição nestas proteínas. Isto foi útil para compreender o comportamento molecular de ligações químicas específicas (por exemplo, ligações de hidrogénio) em proteínas, que também foram analisadas nas conformações de ADN não canónicas (conformações do modelo c e d) e localizadas nas regiões não codificantes. Desta forma, o conhecimento de sistemas moleculares computacionais aplicados a proteínas pode ser usado para construir modelos para as estruturas de ADN encontradas em regiões não codificantes.

O estudo das alterações estruturais em conformações não-B de ADN é muito importante uma vez que, tal como nas proteínas, podem adotar diferentes estruturas relacionadas com propriedades específicas. Além disso, a arquitetura do genoma (codificação versus não-codificação) levou-nos à análise das especificidades da formação de não-B conformações no genoma mitocondrial completo e suas implicações nos processos biológicos (modelo c). Estas estruturas localizadas em regiões não codificantes parecem desempenhar um papel crítico no processo de geração de deleções em moléculas de genoma mitocondrial associadas com determinadas doenças.

Por último, uma nova metodologia para deteção de estruturas de ADN não-B, em regiões não codificantes foi desenvolvido (modelo d). Os dados disponíveis para microssatélites do cromossoma Y foram estudados usando programas computacionais para a previsão de estruturas e novos algoritmos para identificar novas conformações não-B de ADN. A avaliação dessas estruturas foi tentada por meio de simulações de dinâmica molecular, de integração termodinâmica e cálculos MMPBSA (Molecular Mechanics – Poisson Boltzmann Surface Area). As características estruturais moleculares presentes em ADN nuclear (microssatélites) foram inferidas e correlacionadas com diferentes processos biológicos e doenças. Desta análise resultou a previsão da formação de estruturas específicas que podem surgir em ADN de cadeia simples. A ocorrência destas conformações não-B de ADN em regiões não codificantes pode influenciar/regular os processos de transcrição que ocorrem em regiões que codificam proteínas, ou processos que dependem de potencial específico de ‘folding’ como a replicação do ADN.

Há uma clara associação entre as regiões que codificam proteína (modelo a e b) e regiões não-codificantes dos genomas (modelo c e d). A possibilidade de estas duas regiões diferentes, gerarem ou formarem arranjos moleculares tridimensionais foi estudada nesta tese. O ADN não-B pode adotar diferentes conformações, tal como em sistemas de proteínas, o que ficou demonstrado nesta tese. Embora existam características estruturais únicas das proteínas e das estruturas de ADN não-B, os dois diferentes sistemas moleculares podem adotar conformações tridimensionais. Em regiões não codificantes, a formação de conformações de ADN não-B tem implicações na evolução em geral, bem como especificamente em deleções, na etiologia de várias doenças, e na replicação do material genético (modelo c e d).

Table of Contents

2. General Introduction and Discussion List of Figures

3. General Introduction and Discussion List of Tables

4. General Introduction and Discussion

List of Abbreviations

Adenosine diphosphate	ADP
Adenosine triphosphate	ATP
Cytochrome c oxidase	CO
Encyclopedia of DNA Elements	ENCODE
Entropic contribution	TS
Expressed sequence tag	EST
Graphical user's interface	GUI
Hypervariable regions	HVR
Hydrogen bonds	H-bonds
International Union of Pure and Applied Chemistry	IUPAC
Ion mobility spectrometry	IMS
Metallothionein	MT
Messenger Ribonucleic acid	mRNA
Nanoelectrospray mass spectrometry	Nano-ESI-MS
National Center for Biotechnology Information	NCBI
Nicotinamide	Nam
Nicotinamide Adenine Dinucleotide	NAD
Nicotinamide phosphoribosyltransferase	NAMPT
Nicotinic acid	Na
Nucleic Acid Builder	NAB
Nucleic acids database	NDB
Operating system	OS
Poly ADP-ribose polymerases	PARPs
Polymerase chain reaction	PCR
Protein databank	PDB
Research Collaboratory for Structural Bioinformatics	RCSB
Ribonucleic acid	RNA
Ribosomal ribonucleic acid	12S and 16S rRNA
Root-Mean-Square-Deviation	RMSD
Short tandem repeats	STRs
Stepwise mutation model	SMM
Base pairs	Bp
Three-dimensional	3D

Transfer RNAs	tRNAs
Ubiquinolcytochrome c oxidase reductase	Cyt b
Ubiquinone oxidoreductase	NADH
Water	WAT
Watson-Crick	W-C
Molecular Dynamics	MD
Molecular mechanics Poisson-Boltzmann solvent accessible surface area	MMPBSA
Generalized Born	GB
Poisson-Boltzmann	PB
Solvent accessible surface area	SASA
Steered molecular dynamics	SMD

5. General Introduction

5.1. Coding versus Non-coding DNA

Although the molecular structure of DNA has been described long ago [4, 5] (Figure 1), the comprehension of the genome structural architecture is rather difficult due to its sequence plasticity. Processes as DNA replication, recombination, mutation and retro-transposition make difficult to disentangle cause and consequence in DNA structural local or global conformational behaviour. Several approaches have been used to reveal how these processes act and what kind of changes they can produce [6-8]. These studies have shown that several sections of the genome appear to be non-codifying regions without any relevance for living cells and therefore to organisms, with the exception of non-coding functional ribonucleic acid (RNA) and microRNAs. Classically, the non-coding regions represent the section of the genomes where transcription does not occur. Transcription is the first step of gene expression that converts a sequence of DNA is to a specific chemical molecule (messenger ribonucleic acid - mRNA) that then can be converted to protein. There are well organized sections of a gene delimited by the transcribed part (exons) and not transcribed (introns). Only exons are transcribed, but the intronic part has important roles in alternative splicing and other processes [9]. The mRNA->protein conversion, called translation, is a process where each three nucleotides in mRNA represent an amino acid in the protein. The transcribed gene can also represent non-coding molecules (e.g., ribosomal RNAs, micro RNAs). The highly organized architecture of coding regions, where genes are present and transcription occurs, was stated to be not present in large percentage of non-coding regions and that part of genome was called “junk DNA”. Several articles had demonstrated that the so called non-coding regions are indeed relevant and play a role in many cellular mechanisms, from prokaryotes to eukaryotes [10-19]. Recently the Encyclopedia of DNA Elements (ENCODE) project studied transcription, transcription factor association, chromatin structure and histone modification, that revealed biochemical functional regions in almost 80% of the human genome [13-20]. The functional DNA detected does not match protein-coding regions (exome), still it play an important role in the regulation of the genome. The ENCODE database of functional elements might be used to better understand different type of diseases (e.g., cancer, rare genetic disorders, common diseases with a genetic component), and therefore used to elucidate the relationship between functional non-coding DNA and coding DNA. Herein the name non-coding will be used for

regions that are not protein-coding, but where the transcription process can occur, might regulate transcription processes in protein-coding regions (e.g., act like transcription factors), or might present a biochemical signature in the cell. This thesis mainly focus the non-coding regions that are not transcribed (mtDNA control region and STRs), but present a biochemical signature in the cell, and also the importance and relevance of mutation events in the coding regions (MT, NAMPT and PNC genes) to the understanding of non-coding genome.

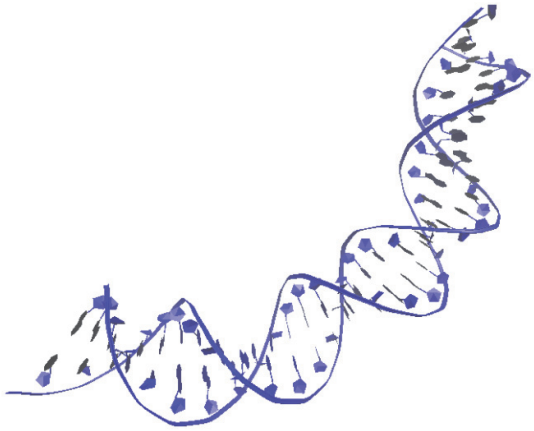


Figure 1: Molecular structure of double-helix DNA showing the base-pairing between nucleotides and backbone conformation (figure generated with VMD[1, 2] software).

The proliferation of “selfish elements” (mainly segments of DNA called mobile elements) until it is prohibitive for the organism survival was suggested to support non-coding regions proliferation, but cannot explain spliceosomal introns, small repetitive DNAs and random insertions[21]. Others defend that non-coding DNA results from natural selection and genome size and should have a direct impact in nuclear volume, cell size and cell division rate [22].

There are approximately 250 full genomes from different prokaryotic species with 350-8000 genes. Eukaryotic genomes are represented by 2455 species already sequenced and the number of genes present in each are higher than 13000 [23-26]. The coding DNA described for prokaryotic and eukaryotic organisms represent only part of the total genome. In Eukaryotes, the genome has a high variation in size but the percentage of coding regions in the genomes remains the same (C-paradox) (Figure 2) [23]. The increase or decrease in genome size results mainly from expansion of introns and mobile elements (non-coding regions). The variation in complexity is possibly explained considering differences in gene deployment: patterns of transcriptional regulation and alternative splicing [23, 27]. The data from ENCODE project corroborates that the differences in gene deployment and transcriptional regulation depend not only on coding segments of the genome, but are associated to non-coding functional elements that interact with gene regions [16, 19, 20].

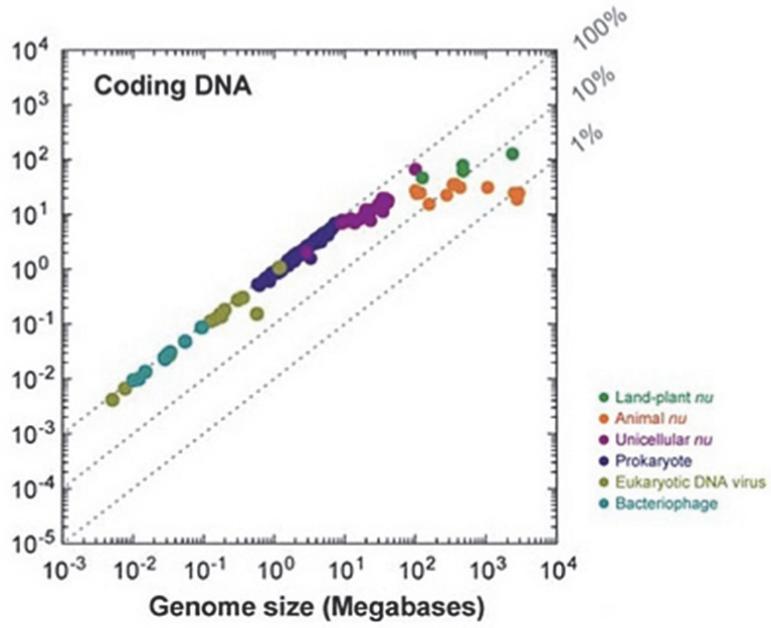


Figure 2: C-value paradox in Eukaryotes: increase of genome size depends of the increase of non-coding elements (figure from Lynch 2007, "The origins of genome architecture").

In humans, coding regions (≈24000 genes) represent 1% of the genome. The non-coding part, which includes non-coding functional RNA cis-regulatory elements, telomeres, introns, pseudogenes, repeat sequences, transposons and viral elements, represents about 99% of the total genomes [13-17, 19, 23]. In this thesis different analysis of some of these non-coding elements (e.g., pseudogenes - model a, control region mtDNA – model c, repeat sequences – model d) and coding regions that can often became non-coding or non-functional (model a, model b), were performed. Different gene regions that are transcribed into proteins were analysed, suggesting that they can become non-functional (by accumulation of critical mutations) in different model species. The relationships between different genomic regions and several biological processes (e.g., gene expression, gene pathways) were also accessed. Regions of genome that are transcribed can become often non-functional by mutational events, or even became non-transcribed elements in the genome (model a, model b). On the other hand, regions that are not subjected to transcription mechanism, therefore not under selective pressures, can have relevant roles associated with structural relevant features of DNA (model c, model d). The DNA regions prone to form any DNA conformation that is not the orthodox right-handed Watson-Crick B-form (non-B DNA) play an important role in critical biological processes (e.g., replication, deletions, transcription) that are now been understood. In this thesis, the analysed structural features of DNA are the non-B DNA conformations (hairpin, cruciform, cloverleaf-like elements and other secondary structures).

We have used four genetic models to address the questions related with non-coding regions: The metallothioneins (model a), NAD pathway relevant genes (model b), MtDNA

(model c), and nuclear short tandem repeats (model d). These models will be described briefly in chapter 5.3.

5.2. Contributions to Articles.

JC contribution to the article related with metallothioneins (model a) was the bioinformatics experiments and computational analysis of the data (e.g., phylogenetic analysis). JC had no participation in the RT-PCR and expressed sequence tag (EST) analyses.

JC contribution to NAD pathway relevant genes article (model b) was the bioinformatics analysis (e.g., protein-ligand binding, calculations of active site interactions). JC did not participate in the laboratory experiments.

In the mtDNA non-B conformations (model c) analysis, JC helped in the development of bioinformatics tools to analyse the data (e.g., python scripts for UNAFold, Circos diagrams) and helped in the interpretation of results. JC had no participation in collecting the data and in the statistical analysis.

JC and ISM designed the experiments and analysed the data for the Y-STRs article (model d). All authors helped in interpretation of results and in writing the article.

JC performed the design of the software and the implementation of algorithms in the SPInDel workbench article. All authors helped to write the article.

5.3. Genetic Models.

5.3.1. Gene Families: The Metallothioneins

Lineage-specific traits and development of novel biological functions may result from pre-existing genes [28-30]. The chance of occurrence of novel biological functions (neofunctionalisation) is expectedly lower than the chance of inactivation (pseudogenisation) [31]. By this way, adaptive changes are less frequent since most amino acid replacements are neutral or deleterious. Models like the mammalian metallothionein family (MT family) can be used to study particular pathways of neofunctionalisation, pseudogenisation or subfunctionalisation [32-34]. Since several genomes (mammalian genomes) are currently available the study MT clusters and the evolutionary steps underlying the expansion of this gene family is possible. MTs are metal-binding proteins involved in homeostasis and the transport of essential metals. They are also relevant in protecting cells against heavy metals toxicity [35, 36], having thus a critical role in many biological processes. The reconstruction of the evolutionary history of MT clusters, combined with the expression profile of MT genes and behaviour of structural interactions of specific residues can help us to understand the relevant features of non-coding regions that result from specific duplications in mammalian genomes.

5.3.2. NAD Pathway Relevant Genes: NAMPT and PNC.

Several redox reactions (chemical reactions in which atoms have their oxidation state changed) occurring in the cells from prokaryotes and Eukaryotes use nicotinamide adenine dinucleotide (NAD) as a cofactor [37-42]. Regulation of metabolism and energy production are mediated by NAD and it can also act as substrate for NAD-consuming enzymes, such as poly (ADP-ribose) polymerases (PARPs) and sirtuins. NAD is involved in DNA repair, transcriptional silencing and cell survival [3].

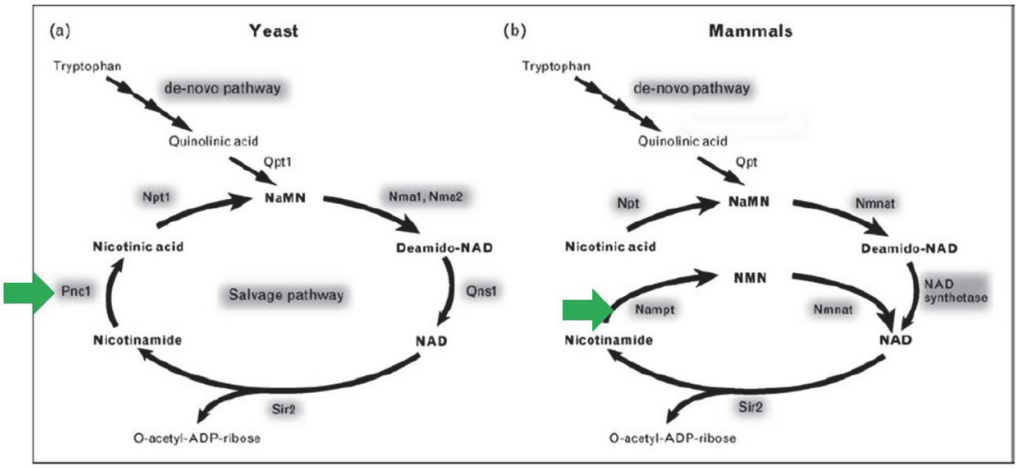


Figure 3: Description of de novo pathways synthesize of NAD from tryptophan or aspartic acid and of the salvage pathways that recycles NAD from nicotinamide (Nam), nicotinic acid (Na) and their ribosides (source figure from Revollo et al. [3]).

The synthesis NAD was studied considering different routes that depend on alternative precursors. *De novo* pathways synthesize NAD from tryptophan or aspartic acid and the salvage pathways recycle NAD from nicotinamide (Nam), nicotinic acid (Na) and their ribosides [39] (Figure 3).

In humans the major source of intracellular NAD results from the nicotinamide salvage pathways [38] but several microorganisms also need this pathway to grow [42-44]. Mammalian cells do not present nicotinamidases which makes them a target to the development of drugs for infectious diseases and anti-parasitic therapies [43-47].

In yeast and invertebrates the nicotinamidase gene PNC1 has been described as biomarker of stress and regulator of sirtuin [37, 48]. There are studies that tried to correlate these enzymes with aging [49] and infection [43-45, 49].

Inflammation and disease has also been associated to the functional homologue of nicotinamidase in vertebrates, nicotinamide phosphoribosyltransferase (*NAMPT*) [50, 51]. Nicotinamidase expression protects human neural cells but increase in PNC1 and sirtuin activity also protects against proteotoxic stress in yeast and *C. elegans* [52, 53].

The two enzymes described before can be present in the same organism [40, 41], rising the question about which one of them is expressed in these species.

5.3.3. MtDNA

MtDNA is a circular stranded molecule with a length of approximately 16.569 base pairs (bp) in humans and is normally present in all animal nucleated cells. It is contained in a double-membrane intracellular organelle (the mitochondrion) that is responsible for the energy generating process of oxidative phosphorylation. The mtDNA usually encodes thirteen important polypeptides in respiratory complexes (NADH - ubiquinone oxidoreductase: NADH1-NADH6 and NADH4L for complex I; Cyt b - ubiquinolcytochrome *c* oxidase reductase for complex III; CO - cytochrome *c* oxidase: COI-III for complex IV; ATP - adenosine triphosphate: ATPase6 and ATPase8 for complex V), two ribosomal ribonucleic acid (12S and 16S rRNA) and twenty two transfer RNAs (tRNAs)[54]. This genome has also a region known as the non-coding region, that is referred as the control region in the literature, with regulatory functions [55]. Two hypervariable regions can be identified (HVRI and HVRII) in the control region.

The human mtDNA has a few unique characteristics, namely a) maternal inheritance [56-58], b) discrete origins of replication, c) intronless genes, e) absence of dispersed repeats, f) few intergenic DNA, f) polycistronic transcripts, g) different genetic code and h) high copy number per cell [55].

Previous studies have determined the importance of several non-B DNA conformations in the mtDNA [59].

5.3.4. Short Tandem Repeats Model: Features and Mutation Mechanism

Short tandem repeats (STRs) represent 3% of human genome [23]. Most are located in non-coding regions. It is assumed that they do not have a biological function so they are classified as “junk DNA”. However there are clues pointing to the influence of STRs in gene expression (e.g., $[CA]_n$ and $[CT]_n$ repeats near a gene), recombination, maintenance of chromatin spatial organization [60, 61]. The mutation rate of these sequences is lower than unique DNA sequences (Table 1) [60].

Table 1: Mutation rates of unique DNA sequences and STR sequences (order of magnitude).

	Mutation rate order of magnitude (nucleotides per generation)
Unique DNA sequences	10^{-9}
STR sequences	10^{-2} to 10^{-6}

Y chromosome STRs are used in most studies (e.g., population genetics, evolution and forensics) as genetic markers [62-67]. The Y chromosome is one of the two sex-determining chromosomes in most mammals, and is a good model to study the contrasts between non-coding and coding regions since there is no recombination, except for the pseudo-autosomal region. The Y-STRs used in our study were retrieved from National Institute of Standards and Technology (NIST) [68] and are described in Table 2 .

Table 2: Allele range, repeat motif, GenBank accession numbers and reference alleles of Y-STR locus. Repeat motif abbreviations A,T,G,C,W,Y,R,S correspond respectively to adenine, thymine, guanine, cytosine, weak (A or T), pyrimidine, purine, strong (G or C) following the International Union of Pure and Applied Chemistry (IUPAC).

Marker Name	Allele Range* (repeat numbers)	Repeat Motif	GenBank Accession	Reference Allele
DYS19	10-19	TAGA	AC017019	15
DYS385 a/b	7-28	GAAA	AC022486	11
DYS389 I	9-17	(TCTG) (TCTA) (TCTG) (TCTA)	AC004617	12
DYS389 II	24-34	(TCTG) (TCTA) (TCTG) (TCTA)	AC004617	29
DYS390	17-28	(TCTA) (TCTG)	AC011289	24
DYS391	6-14	TCTA	AC011302	11
DYS392	6-17	TAT	AC011745	13
DYS393	9-17	AGAT	AC006152	12
YCAII a/b	11-25	CA	AC015978	23
DYS388	10-18	ATT	AC004810	12
DYS425	10-14	TGT	AC095380	10
DYS426	10-12	GTT	AC007034	12
DYS434	9-12	TAAT (CTAT)	AC002992	10
DYS435	9-13	TGGA	AC002992	9
DYS436	9-15	GTT	AC005820	12
DYS437	13-17	TCTA	AC002992	16
DYS438	6-14	TTTTC	AC002531	10
DYS439	9-14	AGAT	AC002992	13
DYS441	12-18	TTCC	AC004474	14
DYS442	10-14	$(TATC)_2(TGTC)_3(TATC)_{12}$	AC004810	17
DYS443	12-17	TTCC	AC007274	13
DYS444	11-15	TAGA	AC007043	14
DYS445	10-13	TTTA	AC009233	12
DYS446	10-18	TCTCT	AC006152	14
DYS447	22-29	TAA**W**A	AC005820	23
DYS448	20-26	AGAGAT	AC025227	22
DYS449	26-36	TTTC	AC051663	29
DYS450	8-11	TTTTA	AC051663	9
DYS452	27-33	**Y**ATAC	AC010137	31
DYS453	9-13	AAAT	AC006157	11
DYS454	10-12	AAAT	AC025731	11
DYS455	8-12	AAAT	AC012068	11
DYS456	13-18	AGAT	AC010106	15
DYS458	13-20	GAAA	AC010902	16
DYS459 a/b	7-10	TAAA	AC010682	9
DYS460 (A7.1)	7-12	ATAG	AC009235	10
DYS461 (A7.2)	8-14	(TAGA) CAGA	AC009235	12
DYS462	8-14	TATG	AC007244	11
DYS463	18-27	AA**R**GG	AC007275	24
DYS464 a/b/c/d	11-20	CCTT	X17354	13
DYS481	20-30	CTT		22
DYS485	10-18	TTA		16
DYS490		TTA	AC019058	12
DYS495	12-18	AAT	AC004474	15
DYS497	13-16	TTA		14
DYS504	11-19	TCCT	AC006157	18
DYS505	9-15	TCCT	AC012078	12
DYS508	8-15	TATC	AC006462	11
DYS520	18-26	ATA**S**	AC007275	20
DYS522	8-17	GATA	AC007247	10
DYS525		TAGA	AC010104	10
DYS531	9-13	AAAT		11
DYS532	9-17	CTTT	AC016991	14
DYS533	9-14	ATCT	AC053516	12
DYS534	10-20	CTTT	AC053516	15
DYS540		TTAT	AC010135	12
DYS549	10-14	GATA	AC010133	13
DYS556		AATA	AC011745	11
DYS557		TTTC	AC007876	16
DYS565	9-14	ATAA	AC010726	12
DYS570	12-23	TTTC	AC012068	17
DYS572	8-12	AAAT		10
DYS573	8-11	TTTA		10
DYS575		AAAT	AC007247	10

DYS576	13-21	AAAG	AC010104	17
DYS594	9-14	AAATA	AC010137	10
DYS607		$[GAAG]_{15}[GAAA][GAAG][GAAA][GAAG]$		19
DYS612		$[CCT]_5[CTT][TCT]_4[CCT][TCT]_{25}$	AC006383	36
DYS626		AAAG		18
DYS632		CATT	AC006371	9
DYS635 (C4)	17-27	T**S**TA compound	AC004772	23
DYS641		TAAA	AC018677	10
DYS643	7-15	CTTTT	AC007007	11
Y-GATA-H4	8-13 (25-30)	TAGA	AC011751	12
Y-GATA-C4	20-25	T**S**TA compound	G42673	21
Y-GATA-A10	13-18	$(TCCA)_2(TATC)_{13}$	AC011751	15

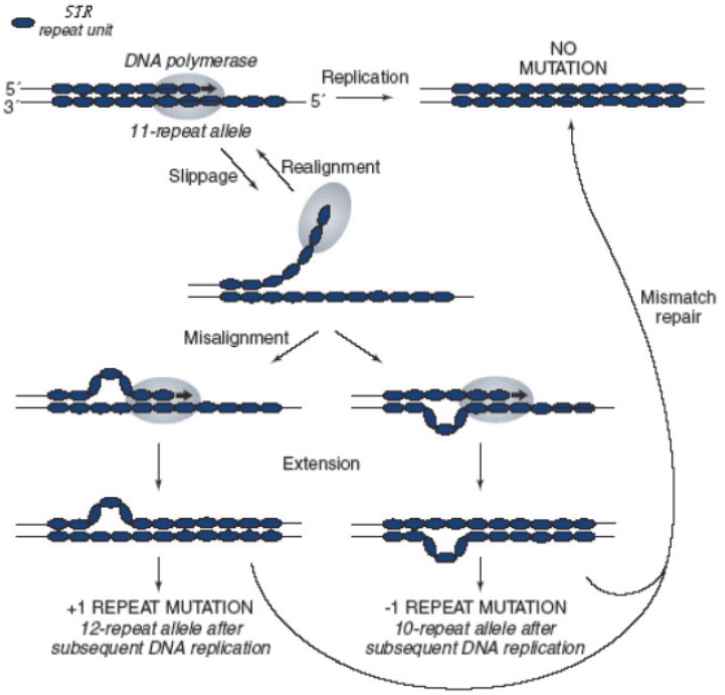


Figure 4: Schematic view of strand-slippage replication mechanism (figure from Jobling et al.[69]).

There are factors that influence in different ways STR mutations such as repeat number, repeat base composition, repeat size, flanking sequence, recombination, sex and age of the individual [60].

STR accurate replication depends of diverse cell machinery that is used during cell division, DNA repair and recombination. DNA polymerases are essential to keep the integrity of the genome at different stages of cell development [70, 71].

One of the models used to explain Y-STR mutation mechanism is the stepwise mutation model (SMM) [69, 72-78]. This model (Figure 4) assumes that only small changes (when assuming that the change is one repeat unit at time we call the model single SMM) in allele number occur, there are equal probabilities of increasing and decreasing of repeat number, the size of alleles is unlimited and there is independence of the rate and size of mutations from the repeat number [60, 69].

The biological mechanism that seems to be involved and can explain the observed results for STR mutations is the strand-slippage replication [69, 79, 80] (Figure 3). This process occurs during replication. After DNA single strand template is generated in 'origins' points that are recognized by proteins (helicases) that separate the two strands, folding of the template or of the copied strand can occur and originate a final DNA fragment with one allele size difference (one step mutation) [69, 79-84]. The techniques that are used to detect differences in STRs allele size are the polymerase chain reaction (PCR) followed by an electrophoresis. The development of the PCR technique has significantly improved the efficiency of laboratorial diagnostic procedures by allowing the *in vitro* formation of a large number of DNA copies (amplification) using a specific genomic region as template [85].

STRs were characterized by different experimental approaches as nanoelectrospray mass spectrometry (nano-ESI-MS) and ion mobility spectrometry (IMS) [100], aside from different *in silico* approaches [101-103]. STR repetitive motifs can interfere in basic molecular mechanisms as DNA replication [70, 79, 104-110].

5.4. Non-B DNA Conformations Prediction

Primary nucleotide sequences are just the tip of the iceberg concerning the role of DNA in cellular processes [10,11, 86-89]. Little attention has been given to other levels of genetic information beyond primary DNA sequences. It has been shown that non-B DNA conformations (any DNA conformation that is not the orthodox right-handed Watson-Crick B-form) can have important roles in DNA replication, transcription and recombination. The existence of conserved structural DNA stretches suggests that such local DNA conformations can be used to estimate phylogenetic relationships. In this regard, several methods have been proposed in the literature for phylogenetic inference from DNA primary sequences [90]. However, these methodologies usually rely in simple genetic distances [91] and/or models of nucleotide substitution [92] disregarding structural DNA information.

By studying evolutionary constrains in secondary and tertiary DNA structures it is possible to have a glimpse of how selective pressures are modulating mutation patterns in DNA and their implications in understanding complex protein–nucleic acid interactions [87, 93]. The study of non-coding DNA structures is facilitated by the large number of nuclear and mitochondrial genomes now available for many species (genomes of National Center for Biotechnology Information database, NCBI at www.ncbi.nlm.nih.gov/sites/entrez?db=genome). The quality of data concerning non-coding regions is improving exponentially, allowing good predictions of structural DNA parameters [10, 86, 87, 93].

It has been shown that a main cause for mutagenic instability is the occurrence of non-B conformations stabilized by negative supercoiling [87]. Large genome rearrangements, deletions or structural polymorphic states could have relevant phenotypic consequences to the organism [86, 87, 94, 95]. For instance, DNA slipped structures play a prominent role in several hereditary neurological diseases (e.g., Friedreich's ataxia, Huntington disease or myotonic dystrophy) and some mtDNA deletions syndromes [86, 87, 94-97]. It has been already described that DNA-binding proteins, phenotype-associated SNPs and predicted enhancers are functionally relevant [10, 98, 99].

There is a lack of studies incorporating structural mutagenic pattern in non-coding genomic regions. Databases as Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB) database and nucleic acids database (NDB) can be used to better understand DNA structural features considering different genomic regions (coding versus non-coding). Specific DNA non-B conformation structures (Hairpin, Pseudoknot, and Cruciform) were associated with errors occurring in replication [100].

5.4.1. Thermodynamics of DNA and UNAFold

Thermodynamics parameters of DNA have been studied for a long time [101-104]. Dynamic programming algorithms for DNA secondary structure prediction were utilized in this thesis. UNAFold [105] is a software that can perform DNA secondary structures predictions for Watson-Crick (W-C) pairings, wobble and non-canonical states under a variety of salt conditions, empirical equations for monovalent and magnesium dependence of thermodynamics. Nearest neighbor energy rules for Watson-Crick base pairs, internal mismatches, terminal mismatches and dangling ends is used to calculate the predicted structures based in a experimentally free energy database that considers different motifs [106]. Energies are also assigned to loops (pseudoknots and base triplets are excluded). The final experimental energy values are assigned to internal, bulge and hairpin loops. Hairpin loop is an unpaired loop in the end of a structure that begins by a paired double helix of DNA. The UNAFold complete database of parameters for base pairs, mismatches, terminal dangling ends, terminal mismatches, coaxial stacking, and a variety of loop motifs including hairpins, bulges, internal loops, and multibranched loops was used. Methods for measurement of the thermodynamic parameters have been reviewed elsewhere [102, 107-110]. Dependence equations are implemented in UNAFold to perform accurate non-B DNA conformations (e.g., secondary structure) calculations for different values of solution conditions, empirical sodium and magnesium.

5.4.2. AmberTools: Molecular Dynamics of Nucleic Acids, NAB and MMPBSA.

AmberTools [111, 112] is a set of tools that can perform different calculations (e.g., build molecular systems, solvate systems, neutralize systems, root-mean-square deviation calculations, end to end distances, hydrogen bonds (H-bonds) calculations, molecular mechanics Poisson-Boltzmann solvent accessible surface area) over three-dimensional (3D) models of proteins or DNA and read molecular simulation data resulting from Amber [111, 112] molecular dynamics calculations.

NAB [113, 114] is a programming language designed to generate models for “unusual" DNA and RNA as the ones predicted by UNAFold. DNA has almost an infinite number of possible conformations with a repeat unit (sugar) that contains seven rotatable bonds (flexible backbone) and a rigid planar base (nucleotide base) [112]. These DNA features difficult the accurate prediction of non-canonical structures (e.g., secondary structures) using refinement methods as molecular mechanics since there are no 3D structures with high homology to our predicted models. Using this high level programming language, residues, strands and molecules can be treated as objects and several routines

can be performed over these objects. Manipulation of axis systems, including rotation and translations, can be implemented with NAB. Different type of models can be built using distance geometry methods with an additional coordinate manipulation for specific constrained systems. Molecular dynamics simulations can be easily implemented using *ab initio* models with the AMBER [111, 112] force field.

The prediction of free energies differences directly linked to conformational equilibria is usually vital to understand the molecular basis of crucial biological functions [115]. The different conformations of DNA that result from properties of the phosphodiester backbone and the nucleic base pairs can be analysed with computational methods, which are also able to determine the associated free energies. Therefore, methods such as molecular mechanics Poisson-Boltzmann solvent accessible surface area (MMPBSA) can be used to determine the free energy of the individual end-point of each DNA molecular system. The entropy and the enthalpy [Generalized Born (GB) and Poisson-Boltzmann (PB)] [116, 117] calculations, must be determined to calculate the contributions to the free energy. Typical contributions to the free energy include the internal energy (bond, dihedral, and angle), the electrostatic and the van der Waals interactions, the free energy of polar solvation, the free energy of nonpolar solvation, and the entropic contribution (TS):

G_molecule= E_internal+E_electrostatic+ E_vdW+ G_(polar solvation)+ G_(non-polar solvation)- TS (1)

For the calculations of relative free energies between closely related complexes, it is assumed that the total entropic term in equation 1 is negligible as the partial contributions essentially cancel each other [118]. The first three terms of equation 1 can be calculated with no cut-off. The nonpolar contribution to the solvation free energy due to van der Waals interactions between the solute and the solvent is usually modeled as a term dependent of the solvent accessible surface area (SASA) of the molecule [119].

5.5. Python Programming in DNA Analysis

A wide range of computer programs is now available to deal with the huge amount of genetic information generated in thousands of laboratories around the world. The appropriate choice of a program for a given task depends both on the data and on the goals of the experiment. For instance, many open and closed source programs are available to make phylogenetic and evolutionary inferences from genetic data [120, 121].

The first step to build a computer program for management and analysis of genetic information is to choose the appropriate programming language (e.g., Python, Java, Perl, and C++). Another important aspect that must be considered in a software development effort is the interface with the user. To build an easy-to-use program based on point-and-click action over windows buttons, an appropriated graphical user's interface (GUI) must be developed [122]. Conversely, the GUI development is not needed if the main users of the program are familiarized with commands via prompt operating system (OS) console window. Thus, very important issues for the main core of a program are: the data to be analysed (input data), the implementation of algorithms to perform the calculations or simulations over the data and the format of final results (output data). Different input file formats are normally used to store molecular data like DNA or proteins, namely Phylip [123], FASTA, MEGA [124], NEXUS, GenBank and protein databank (PDB). These formats are commonly used as input or output formats in several programs (e.g., Phylip, MEGA, PAML[125], MrBayes [126], DnaSP, Bioedit [127]) and standard molecular databases (e.g., GeneBank, FASTA, eXtensible Markup Language-XML). Conversion between different input file formats is usually possible and extremely useful if the user wants to carry out different kind of analyses over the data.

5.5.1. Python

Python (free available at www.python.org) is an object oriented language created by Guido van Rossum [128] that has gained attention in recent years. As other high programming language it can only be executed after processed by a computer. Although being slower than low programming languages, Python have some important advantages: a) a reduced programming time, b) a shorter and easier to read source code c) a high productivity and d) a multiplatform capability (Windows, Linux, and Mac). Different Python third-part modules can be installed for a large variety of tasks, including molecular data handling and analysis: BioPython[129], PyCogent [130] , Matplotlib, , GenomeDiagram [131], NetworkX, py2exe, NumPy , Psyco, SciPy , WxPython (Table 3). These packages have the same common terminology of Python language although with specific modules and built-in functions. An extensive documentation comes as part of Python distribution [128] or can be found in dedicated books and articles [132, 133].

Table 3: Relevant Python modules in molecular data analyses.

Module	Functionality	Requirements	Documentation	Major flaws
BioPython	-Parse bioinformatic files into Python for several formats -Management and manipulation of genetic and proteic data -Code to perform searches in common on-line bioinformatics databases destinations (e.g., NCBI)	-Python 2.3 or later -Numerical Python	-Good and well written documentation	-Some bugs resulting for poor maintenance of some functions
PyCogent	- Same as BioPython	-Python 2.4 or later	-Good and well written documentation	-Some bugs
GenomeDiagram	-Graphic representation of genomes and DNA sequences	-Python 2.4 or later	-Good documentation	
Pythia	-Thermodynamic calculations	-Python 2.4 or later	-Bad documentation	Some bugs
Matplotlib	-Plot and save graphics in different formats -Handle geographic maps	-Python 2.4 or later -Numpy 1.1 -Libpng 1.1 -Freetype 1.4 -Basemap 0.99.2	-Extensive documentation -Very good examples	-Some problems with integration with other major modules, namely wxPython
NetworkX	-Construct phylogenetic relationships through networks design and visualization	-Python 2.4 or later	-Documentation not enough -Lack of good examples	Bugs and lack of flexibility related with visualization and drawing
NumPy	-N-dimensional array object -Linear algebra functions -Basic Fourier transforms	-Python 2.4 or later	-Nice and exhaustive documentation	-
Psyco	-Speed up the execution of any Python code	-Python 2.4 or later	-Bad documentation	-Maintenance and updates very limited
Py2exe	-Converts Python scripts into executable Windows programs able to run without requiring a Python installation	-Python 2.3 or later	-Good documentation	-Poor stability of executables
SciPy	-Language extension that uses numpy to do advanced math, signal processing, optimization, statistics	-Python 2.4 or later	-Very well organized documentation -A lot of cookbook examples	-
wxPython	- Allows easy creation of robust, highly functional graphical user interface	-Python 2.3 or later	-Good wxPython reference documentation -wxPython demo with examples for the code	-Slow performance -Does not include a rapid application development tool (RAD)

5.5.2. Python Language and WxPython: Code and Common Terminology

As an object-oriented language the functionality of Python is based in objects. Objects can be primitive data (integer, float, Boolean and complex), collection data (string, list, tuple, set dictionary) or even more complex data structures (e.g., SQL databases) [128]. In Python language almost everything can be an object. Normally, a routine process in Python is compacted in a module (a Python file or files saved in plain text with extension *.py) that contains executable statements as well as definition of functions, classes and methods. These Python files are normally edited in an integrated development environment, such as IDE (e.g., VisualWX; http://visualwx.altervista.org/), Eclipse (http://www.eclipse.org/) or NetBeans (http://www.netbeans.org/). IDEs permit that common statements and built-in functions are easily identified in code. Another relevant feature of Python is the mandatory indentation that makes easier to read and write the code.

WxPython is an interface for the C++ toolkit wxWidgets. Cross-platform applications can be created with the functionality of C++ Widgets and the simplicity of Python language [122]. The range of possibilities to build a GUI might be increased by additional widgets (Table 4) that are directly written in wxPython.

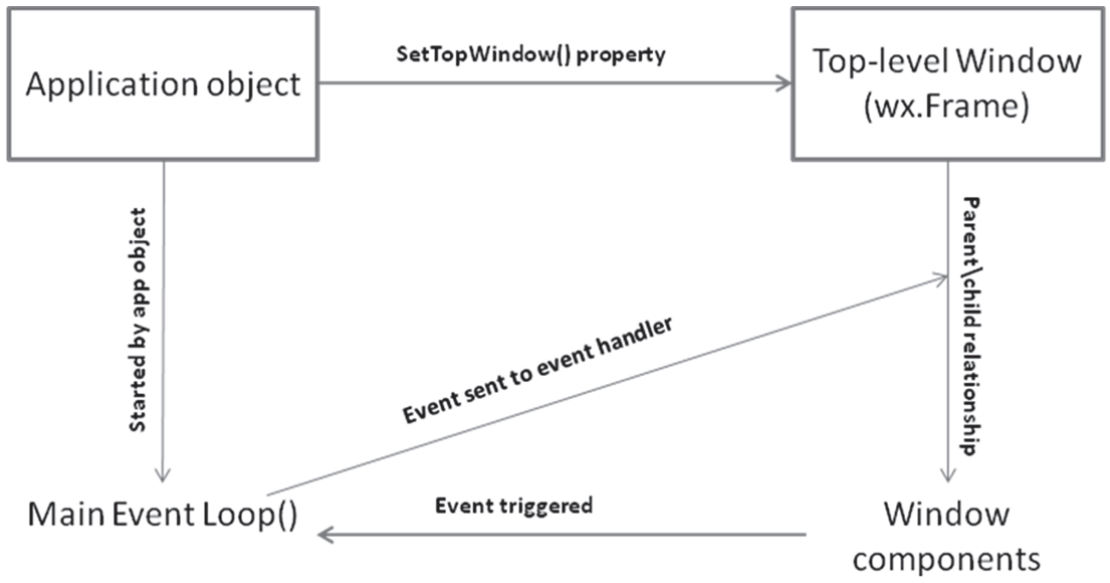


Figure 5: Basic wxPython application structure [adapted from wxPython in Action [122].

Table 4: Common widgets and dialogs implemented in wxPython.

Name	Features	Type
wx.Window	WxWindow is the base class for all windows and represents any visible object on screen. It includes controls and top level windows.	Frame
wx.FlexGridSizer	Lays out its children in a two-dimensional table	Sizer
wx.StaticBoxSizer	Rectangle drawn around other panel items to denote a logical grouping of items	Sizer
wx.Button	Control that contains a text string	Widget
wx.ComboBox	Displays static list with editable or read-only text field; or a drop-down list with text field; or a drop-down list without a text field.	Widget
wx.Grid	WxGrid and its related classes are used for displaying and editing tabular data. They provide a rich set of features for display, editing, and interacting with a variety of data sources, namely genetic data.	Widget
wx.Notebook	Manages multiple windows with associated tabs	Widget
wx.StaticText	Displays one or more lines of read-only text	Widget
wx.TextCtrl	A text control allows text to be displayed and edited; it may be single line or multi-line.	Widget
wx.FileDialog	File chooser dialog	Dialog
wx.MessageDialog	Dialog that shows a single or multi-line message, with a choice of OK, Yes, No and Cancel buttons.	Dialog
wx.ProgressDialog	Dialog that shows a short message and a progress bar	Dialog
wx.SingleChoiceDialog	Shows a list of strings and allows the user to select one	Dialog

5.5.3. BioPython, PyCogent, GenomeDiagram, and Pythia

As described before, Python and wxPython are packages that implement in the main core of the software the routine operations and tools for GUI construction, respectively. It is then necessary to have a set of packages that could easily deal with most common operations required in molecular data manipulation. BioPython is a package that consists in a set of modules to read and manipulate molecular data (DNA and proteins). The most relevant functionalities of BioPython for computational molecular biology are: a) the capacity for parsing bioinformatic files into Python from several formats (Blast out, ClustalW, FASTA, Genbank, PubMed and Medline, Expasy, SCOP, UniGene, SwissProt); b) the incorporation of a code to perform searches in common on-line bioinformatics destinations (NCBI, Expasy); c) the easy management of sequence features (sequence translation, transcription, weight calculation, alignments) and e) the easy integration with BioPerl and BioJava modules through BioCorba [129]. To use DNA and proteins sequences as input data, it is not necessary to write the code since BioPython already has the SeqIO system that defines SeqRecord objects to manipulate this data and normally is very fast reading and manipulating sequences. BioPython module can be called by using ‘import BioPython’ in the beginning of Python file and specific functions are invoked using ‘from BioPython import ‘function’ ’. Documentation for BioPython has many useful examples for all functions. PyCogent [130] is a python module that can perform all the functions implemented in BioPython but focused in genomic biology. GenomeDiagram [131] can make graphic representations of genomic data. Pythia (http://sourceforge.net/projects/pythia/) includes modules that can calculate DNA binding and folding energies of specific DNA sequences.

5.5.4. Matplotlib, NumPy and SciPy

The implementation of mathematical routines when developing software can be achieved by using a great number of external Python modules, although the most commonly used in software development are Matplotlib, NumPy and SciPy. With these modules it is possible to call mathematical functions, to represent the data graphically, to perform iteration over different numerical data and statistical analyses of data. Global statistics from a sequence or alignment, namely proportion of nucleotides and GC content can be calculated with this module.

5.5.5. SPInDel Workbench

The SPInDel workbench is a computational platform developed in python object oriented language using BioPython (http://biopython.org/), SciPy (http://www.scipy.org/), GenomeDiagram (http://bioinf.scri.ac.uk/lp/programs.php), Matplotlib (http://matplotlib.sourceforge.net/), NumPy (http://numpy.scipy.org/), and PyCogent(http://pycogent.sourceforge.net/). It can import alignments of specific targeted genome regions (e.g., *ribosomal RNA* gene regions) showing regions of nucleotide conservation and variation. The variation introduces gaps in the alignment (-) that can be used as a source of information to characterize and classify different species. The classification of each species is based on the different length of the sequence in the alignment that results from insertion/deletion (indel) events.

Theoretically, the discrimination of all Eukaryotic species on Earth (5-15 million) can be done using 6 hypervariable regions with 20 alleles each. The SPInDel analysis was based in *ribosomal RNA* gene regions but the analyses of other regions with the same pattern of sequence evolution (e.g., non-coding regions) is also possible. Different statistical approaches were implemented in this multi-platform software (Windows, Linux or compilation in other operating systems using the SPInDel Workbench source code) by using python algorithms and modules. The application of this software can be extended to other fields where the identification of species is relevant (e.g., ecology, forensics).

5.5.6. NABpy

NABpy (NAB python implementation) is a python module that automatizes all the processes related with initial protein and DNA three-dimensional molecular systems (in vacuum or with explicit solvation) using Matplotlib, BioPython, PyCogent, UNAFold and AmberTools. Different functions are implemented in the module:

- Create protein and DNA molecular systems taking in consideration specific unconstrained and constrained models.
- Solvate systems with explicit water (WAT) and neutralization with sodium ions (Na^+).
- Generate input files to run AMBER [112] molecular dynamics (MD) simulation, including the *.prmtop (topology file) and *.mdcrd (simulation parameters file).
- Calculate H-bonds along trajectories and calculate parameters for DNA base-stacking.
- Calculate the end-to-end distances of all atoms and backbone atoms
- Run structural analysis of DNA molecular systems using Curves+ [134] and 3XDNA [135-137] (helical and backbone parameters).
- Calculate free energy parameters using MMPBSA [112] and Delphi [138].
- Generate graphic representations of results using Matplotlib (e.g., RMSD values).

6. Research Questions and Objectives

The main objective of this thesis is to study non-coding DNA regions in comparison with the already well-studied protein-coding regions and thus to infer which biological processes occur in non-coding genomic tracts of living cells. Using four different research models, the specific objectives of this work were:

- Analyse the MT clusters duplicated genes considering their coding/non-coding status (model a).
- Study NAMPT and PNC genes and respective functional proteins involved in NAD pathways, using different model species (model b).
- Access the current functional status of NAMPT and PNC homologues genes, using a computational methodology with model organisms (model b).
- Perform a protein-ligand docking using homology modelling structures of NAMPT and PNC (model b).
- Detect and identify conserved structural patterns (non-B DNA conformations) in non-coding DNA regions of mammalian mitochondrial (model c) and nuclear genomes (model d).
- Evaluate the association between conserved non-B DNA conformations and specific types of non-coding regions such as mitochondrial control regions (model c) and STRs (model d).
- Measure the degree of randomness of structural conservation across genomes using statistical methodologies to validate identified structures. Infer evolutionary constrains and mutagenic patterns in identified structures (model c, model d).
- Identify secondary structures in mtDNA and their role in different biological processes (model c).
- Determine the structural features of different regions in mtDNA (coding and non-coding), and ascertained how these non-B DNA conformation might influence genetic disorders, replication and transcription (model c).
- Implement a 3D structural analysis of DNA using python algorithms, UNAFold, and non-B DNA conformations database (model d).
- Structural analysis of DNA using previously described computational methodologies, such as molecular dynamics (model d).
- Correlate the size, localization and physical parameters of predicted structures with specific genomic features: replication origins, transcription and mutagenic instability (model d).
- Design software able to use different regions of the genome (e.g., non-coding regions), in order to identify taxonomic groups at various levels (SPInDel workbench).

- Design software (NABpy) to automatize DNA molecular dynamics simulation and MMPBSA free energies analysis.

Next chapters reflect the work that has been achieved in order to tackle the research questions and objectives focussed upon during this work.

7. Publication I: Gains, Losses and Changes of Function after Gene Duplication: Study of the Metallothionein Family

Gains, Losses and Changes of Function after Gene Duplication: Study of the Metallothionein Family

Ana Moleirinho[1], João Carneiro[1,2], Rune Matthiesen[1], Raquel M. Silva[1], António Amorim[1,2], Luísa Azevedo[1]*

[1] IPATIMUP - Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal
[2] Faculty of Sciences of the University of Porto, Porto, Portugal
*E-mail: lazevedo@ipatimup.pt

Abstract

Metallothioneins (MT) are small proteins involved in heavy metal detoxification and protection against oxidative stress and cancer. The mammalian MT family originated through a series of duplication events which generated four major genes (MT1 to MT4). MT1 and MT2 encode for ubiquitous proteins, while MT3 and MT4 evolved to accomplish specific roles in brain and epithelium, respectively. Herein, phylogenetic, transcriptional and polymorphic analyses are carried out to expose gains, losses and diversification of functions that characterize the evolutionary history of the MT family. The phylogenetic analyses show that all four major genes originated through a single duplication event prior to the radiation of mammals. Further expansion of the MT1 gene has occurred in the primate lineage reaching in humans a total of 13 paralogs, five of which are pseudogenes. In humans, the reading frame of all five MT1 pseudogenes is reconstructed by sequence homology with a functional duplicate revealing that loss of invariant cysteines is the most frequent event accounting for pseudogenisation. Expression analyses based on EST counts and RT-PCR experiments show that, as for MT1 and MT2, human MT3 is also ubiquitously expressed while MT4 transcripts are present in brain, testes, esophagus and mainly in thymus. Polymorphic variation reveals two deleterious mutations (Cys30Tyr and Arg31Trp) in MT4 with frequencies reaching about 30% in African and Asian populations suggesting the gene is inactive in some individuals and physiological compensation for its loss must arise from a functional equivalent. Altogether our findings provide novel data on the evolution and diversification of MT gene duplicates, a valuable resource for understanding the vast set of biological processes in which these proteins are involved.

PLoS ONE 6(4): e18487. doi:10.1371/journal.pone.0018487
Editor: Vincent Laudet, Ecole Normale Supérieure de Lyon, France
Received October 9, 2010; Accepted March 8, 2011; Published April 25, 2011

Introduction

When a particular gene is constrained to a specific function, the appearance of biological novelty demands genetic redundancy. Duplication of pre-existing genes may lead to the establishment of lineage-specific traits and to the development of novel biological functions [1,2,3,4,5]. However, the probability of widening biological functions (neofunctionalisation) is expectedly lower than the chance of inactivation (pseudogenisation) [6,7,8] as most amino acid replacements are more likely neutral or deleterious, rather than leading to any particular adaptive change. Although the majority of gene duplicates result in pseudogenes, many remain functionally active longer than it would be expected by chance. This observation led to the development of the subfunctionalisation model [9,10], according to which the accumulation of complementary loss-of- function mutations within regulatory segments of both members would facilitate their preservation while maintaining the original function. In case of preserving the parental function, duplicates may act as backup compensation copies to buffer against the loss of a functionally related gene [11,12].

The current availability of several genome sequences allows the study of the evolutionary steps underlying the expansion of a gene family by detailed characterisation of lineage-specific expansions. MTs are metal-binding proteins involved in homeostasis and the transport of essential metals, more specifically, in protecting cells against heavy metals toxicity [13,14], having thus a critical role in many biological processes. In mammals, four tandemly clustered genes (MT1 to MT4) are known. Although all genes encode for conserved peptide chains that retain 20 invariant metal-binding cysteines, MT3 and MT4 seem to have developed additional properties relatively to MT1 and MT2, such as protection against brain injuries [15,16] and epithelial differentiation [17], respectively. Finally, during the evolution of the lineage that led to modern humans, MT1 has undergone further duplication events that have resulted in 13 younger duplicate isoforms [18]. The co-existence of younger and older duplicates is thus an opportunity to reconstruct the evolutionary history behind the divergence of the MT family in mammals.

Materials and Methods

Phylogenetic analyses

Coding sequences annotated as orthologues of the human MT genes were extracted from the Ensembl database (www.ensembl.org, release 56: Sep 2009) [19]. The final set of sequences (Table S1) does not include shortened sequences and those annotated in non-human species as representing the orthologue of distinct human MT1 genes. Codon sequences were aligned using MUSCLE [20,21] incorporated in Geneious software v5.1.3 (http://www.geneious.com). Coding MT sequences from four fish species (*Danio rerio*, *Oryzias latipes*, *Tetraodon nigroviridis* and *Takifugu rubripes*), two birds (*Gallus gallus* and *Taeniopygia guttata*) and a reptile (*Anolis carolinensis*) were used to outgroup the phylogeny. Two methods were used to reconstruct the tree topology: maximum likelihood (ML) and Bayesian. In both cases, the model of nucleotide substitution used was HKY+G as determined in jModelTest [22]. The program BEAST [23] was used to estimate the Bayesian phylogeny in two runs (50 million generations each) using a Bioportal at the University of Oslo (http://www.bioportal. uio.no). The resulting log file was analyzed in Tracer [24]. The tree was obtained in TreeAnnotator from the BEAST software using a threshold for clade credibility of 0.5. For all the statistics obtained, the effective sample size (ESS) was always within the recommended threshold. The ML topology (Figure S1) was obtained with PHYML (http://www.bioportal.uio.no) [25] using the transition/transversion ratio, the proportion of invariable sites and the gamma parameter estimated by the program. Bootstrap branch support was estimated using 1000 data sets. Tree visualization and final edition were performed in FigTree v1.3.1 (http://tree.bio.ed.ac.uk/software/figtree).

Organization of the human and mouse MT family

The chromosomal organization of the MT family and flanking neighbours (BBS2 and NUP93) in humans and mice was performed using NCBI (*Homo sapiens* build 36.2 and *Mus musculus* build 37.1) [26] and Ensembl (release 56) genomic coordinates.

RT-PCR and expressed sequence tag (EST) analyses

Total RNA from 15 human tissues was obtained from Ambion (FirstChoice Human Total RNA Survey Panel). The complementary DNA (cDNA) was synthesized with random hexamer primers from 2 mg of human total RNA using the RETROscrip First Strand Synthesis Kit (Ambion) according to the manufacturer's instructions. PCR primers used for amplification of the MT2, MT3 and MT4 transcripts were: 5'ATCCCAACTGCTCCTGCGCCG3' (forward) and 5'CAGCAGCTGCACTTGTCCGACG3' (reverse), 5'CTGAGACCTGCCCCTGCCCTT3' (forward) and 5'TGCTTCTGCCTCAGCTGCCTCT3' (reverse) and, 5'CCCCAGGGAATGTGTCTGCATGT3' (forward) and 5'GGCACATTTGGCACAGCCCGG3' (reverse), respectively. Samples were amplified with Qiagen Master Mix for 35 cycles at 95°C for 30 sec, 62°C for 30 sec and 72uC for 45 sec after an initial denaturation at 95°C for 15 min and followed by a final extension step of 10 min at 72°C. PCR products were then purified with ExoSAP-IT (USB Corporation, Ohio, USA) by incubation at 37°C for 15 min, followed by enzyme inactivation for 15 min at 85°C. The resulting purified fragments were sequenced using an ABI Big Dye Terminator Cycle Sequencing Ready Reaction kit (Applied Biosystems) and analysed in an ABI PRISM 3130xl (Applied Biosystems) for validation of the corresponding sequence.

ESTs were extracted from Unigene [26] as counts per million transcripts for each of the given tissues and displayed as $\log_2$ of transcripts per million. A unique EST for human MT4 is annotated in Unigene (GenBank BF759140.1) but since it contains intronic sequence was removed before plotting the data set. The heat-maps were constructed with an in-house tool using average linkage and the correlation of expression patterns as a measure of dissimilarity to build the hierarchical clusters.

Polymorphic population data

The Biomart tool [27] at the Ensembl database was used to assess the human polymorphic variability. Allelic and genotypic frequencies of Tyr30 (rs666636) and Trp31 (rs666647) were extracted from the HapMap project (http://www.hapmap.org). Genotypes for each population were also retrieved and analysed to evaluate the linkage disequilibrium between variants.

Structural homology modelling and visualization

The crystal structure of the mouse Mt2 [28] (PDB code 4mt2) was used to model [29] both human native and mutation-carrying MT4 proteins following a previously reported methodology [30]. All the structures were displayed with PyMol (www.pymol.org).

Results

Evolutionary history of the MT cluster in mammals

The phylogeny of the MT family in mammals was reconstructed with Bayesian and maximum likelihood methods both approaches resulted in a similar topology (Fig. 1 and Figure S1) revealing the robustness of the inference. Tree topology, supported by high values of posterior probabilities (Fig. 1), point clearly to two rounds of duplication occurring at the MT family creating MT4 first and the ancestor of MT1/MT2/MT3. In a second round of duplication, MT3 diverged from the MT1/MT2 ancestor. These data are in agreement with previous observations [31] which concluded that MT duplication occurred before mammalian radiation. However, because MT genes from birds and reptiles group within the MT1/MT2/MT3 cluster, it is possible to consider a more ancient origin for MT4, and therefore, its loss in non-mammalian land vertebrates. An alternative explanation is that MT4 is thus a mammalian-specific gene which has been accumulating a pronounced number of replacements assigning its sequence to a basal position in both Bayesian and ML phylogenies.

Extant mammalian genomes seem to carry a single copy of MT2, MT3 and MT4, while several MT1 copies are found in some species. The highest number of MT1 genes was found in the genome of primates indicating they have arisen in recent duplication events. The detailed genomic organization of the MT cluster in humans and mice provides a good example (Fig. 2). In both species, genes are oriented as cent-MT4-MT3-MT2-MT1- tel and flanked by BBS2 and NUP93, revealing an evolutionarily conserved arrangement of the cluster. Still, a marked difference distinguishes both genomes. While mt1 did not expand in mice, humans carry 13 arrayed duplicates (MT1A to MT1J, MT1L, MT1M and MT1X), five of which (MT1L, MT1J, MT1D, MT1C and MT1I) have been predicted to be no longer active forms. Overall, this corresponds to a genomic expansion of about 66 Kb in humans.

Characterisation and divergence of mammalian MT proteins

Mammalian MTs are proteins with 20 invariant cysteines (Fig. 3A) which are responsible for the binding and sequestration of zinc (Zn), cadmium (Cd) and copper (Cu), among other metals. A total of 9 and 11 cysteines are required to form protein-domains that bind three and four ions (Fig. 3B). As documented previously [32,33] and illustrated here (Fig. 3B), these residues are involved in metal binding through their thiol (-SH) moieties, which must be oriented towards the inside of the metal clusters whenever the ions are sequestered [34]. In humans and mice, MT1 and MT2 are 61- residue proteins (Fig. 3A). The MT3 holds an extra residue in the b domain (Thr5), which is important for its neuroinhibitory activity [35,36], and a six-residue long insertion in the a domain, whereas MT4 shares an additional residue in the b domain (Glu5) (Fig. 3A).

In order to obtain a clear picture of the amino acid conservation between all pairs of active MTs, identity scores were calculated in human/mouse comparisons (Fig. 4). As shown, MT2, MT3 and MT4 orthologues share 86%, 87% and 94% of residue identity, respectively. Among human MT1 duplicates, MT1E showed the highest identity (85%) with the mouse Mt1. Protein identity scores for MT3 resembles that observed in MT1 and MT2 while MT4 orthologues are strongly conserved in their amino acid sequences (94% of residue identity between human and mouse). Previously, it was suggested that the preservation of MT4 sequence (only 4 out of 62 residues differed between human and mouse proteins) results from functional constraints [37] involved in epithelial cell differentiation [17]. Phylogenetic analyses (Fig. 1 and Figure S1) show that MT4 resulted from an old event of duplication and the high degree of sequence conservation between human and mouse orthologues strongly points to a role which seems to have been functionally important in mammalian evolution.

About the pseudogenisation incidents

An important aspect related to the fate of a gene copy is the identification of the type of replacements leading to the pseudogenisation of functionally active genes. From its genomic sequence, we would be able to reconstruct the ancestral functional open reading frame and, at the same time, discern the panel of mutational events that have accumulated over time. To reconstruct the open reading frame of all the five MT1 pseudogenes, the corresponding genomic sequences were aligned with each of the MT1E exons separately, followed by manual inspection of sequence homology (Fig. 5). Because any attempt to disclose the impact of nonsynonymous replacements is not straightforward unless complemented with additional functional assays, we focused our attention on obvious damaging mutations, such as (a) replacement of invariant cysteine residues, (b) introduction of aromatic residues, as well as (c) premature stop codons, indels and mutations at the consensus donor (GT) or acceptor (AG) splice sites. Because these are shared features among all functional genes, our inferences are not constrained by the functional template that could have been chosen to infer the reading frame of each pseudogene.

We detected a total of 12 deleterious mutations within the predicted reading frame of MT1 pseudogenes, eight of which would result in the cysteine replacement and, in some cases, the inclusion of a premature stop codon (Cys5Tyr, Cys15Tyr, Cys24Tyr, Cys26X, Cys37Tyr, Cys41X, Cys50X, and Cys60Arg) covering MT1JP, MT1CP, MT1DP and MT1LP, three non- cysteine codons that would encode for an aromatic residue (Gly11Phe, Ser18Phe and Ser35Phe) in MT1IP and MT1CP, and a 1-bp deletion at the C-terminal domain in MT1IP. Since cysteine replacement either with tyrosine or with a stop codon involves only a single nucleotide substitution, these residues are expected to often contribute to MT pseudogenisation.

Outside the coding sequence, two mutations were found at the donor and acceptor splice site of MT1CP and MT1IP, respectively. Regarding the number of mutations, MT1CP is the pseudogene harbouring the highest number of events (six in total), followed by MT1IP and MT1DP (with three mutations each), MT1JP (two mutations) and finally by MT1LP, which would encode a truncated protein due to a premature stop codon.

Expression profile of the MT genes in humans and mice

Since gene duplication often results in a diversified spatiotemporal pattern of expression of duplicate family members [38,39,40,41,42] we next examined the MT transcription pattern using EST data for 30 human and mouse tissues as a metric of basal expression for all functional genes. Previous data have shown that MT1 and MT2 are ubiquitously expressed in both species whereas MT3 and MT4 present a confined expression in human and mouse brain [43] and in mouse epithelial tissue [17] respectively,

although data regarding expression of MT4 in human tissues is still missing in the literature. The EST records were assembled in a diagram (Fig. 6A) that, in general terms, strongly overlaps the literature data. For instance, mouse mt1 and mt2 seem to be as widely expressed as the human MT2, a gene that has been proposed to have a housekeeping role for heavy metal homeostasis in every cell [44]. In humans, the expression pattern of MT2 clusters with MT1E and MT1X, the two duplicates that reveal the most wide pattern of basal expression [45]. The remaining genes revealed a more confined pattern of basal expression. For instance, constitutive expression of MT1B seems to be restricted to connective and blood tissues, whereas MT1A is highly expressed in the intestine and adipose tissue and less in uterus, eye, liver and lung. This analysis also rank MT1H, MT1F and MT1G in an intermediate position, showing a pattern that is not as wide as that of MT1E and MT1X but not so restricted as that of MTB, MT1A and MT1M either. Although the lack of ESTs in particular tissues may indicate difficulties to distinguish between different MT1 transcripts or bias in tissue representation, these caveats do not necessarily challenge the tissue-specificity generally observed in most MT1 duplicates, possibly playing distinct roles in distinct cell types as was suggested before [44,46].

In contrast to MT1 and MT2, both MT3 and MT4 are constitutive tissue-specific isoforms that do not respond to metal- induction [17,44,47,48,49]. The EST data (Fig. 6) although supporting high expression of MT3 in human and mouse brain tissues also reveal abundant expression in other tissues as well.

Concerning the MT4 profile, EST records agree with previous findings that documented the stratified squamous epithelium of digestive and reproductive systems as the main source of gene transcripts in mouse [17]. Thus far, no data exist on the expression of MT4 in humans (reviewed in [31]) and the EST surveillance has not detected expression in any tissue. These intriguing observations directed a more refined analysis concerning MT4 expression in humans that was here accomplished by RT-PCR in 15 human tissues. For comparative purposes we also assayed the expression of MT2 and MT3 in the same tissue collection (Fig. 6B). As expected, transcripts related with MT2 were observed in all tissues analysed. A similar scenario was observed for MT3 which is here demonstrated to be as ubiquitously expressed as is MT2. On the other hand, the detection of MT4 transcripts was only possible in four tissues (brain, testes, thymus and esophagus), with an evident higher expression in thymus, whose medullar component is mainly composed of epithelial cells [50]. Although the RT-PCR itself cannot provide exact quantitative inferences, it is possible to infer that in all the tissues where MT4 transcripts were detected, the expression level resulted lower than that of MT2 and MT3 with the exception of thymus.

Human polymorphic variability at the MT cluster

Although it is well established that functional MTs preserve a sequence with 20 invariants cysteines, no studies have thus far established the polymorphic status of the remaining residues. To achieve such information, we retrieved the available information on all human active proteins regarding nonsynonymous replacements (Table 1). Several nonsynonymous variants were found in MT1 duplicates and in MT2, but none is documented for MT3. Although a link between any of these replacements and a functional perturbation is not straightforward without complementary experimental assays, it should be mentioned that, in most of the cases the conserved cysteine residues remain unchanged and the replacement does not introduce an aromatic amino acid. The exceptions were observed in MT1B and MT4. In the MT1B case, a putative deleterious Cys19Ser (rs61744104) is indicated as polymorphic in humans although no additional data concerning the allelic frequencies are available in the databases. In the MT4 case, two candidate deleterious mutations were detected, Cys30Tyr (rs666636) and Arg31Trp (rs666647), both of which result in the introduction of aromatic amino acids along with the substitution of a critical cysteine at position 30. Taking into account the functional requisites of MTs, any of these mutations would ultimately result in an impaired metal-binding protein. In contrast with the MT1B case, allelic frequencies for these two polymorphisms are available and were retrieved from the HapMap (Table 2). Allelic frequencies of Cys30Tyr (rs666636) and Arg31Trp (rs666647) in 11 populations are presented in Table 2. In European populations up to 8,5% of the individuals carry a putatively deleterious allele. Frequencies are even higher in African and Asian populations (up to 30%). Genotype evaluation in all the populations showed that every chromosome that contains the Trp31 allele also contains the Tyr30, but not vice versa, which points to Cys30Tyr as the oldest variant and Arg31Trp as appearing afterwards in the same background.

Discussion

The history and fate of post-duplication events in the mammalian evolution was herein explored by the study of MT family members. After a duplication event, newly arisen genes can follow distinct evolutionary paths: if the parental gene is maintained active, redundant duplicates can escape purifying selection and start to accumulate loss-of-function mutations resulting in pseudogenisation; less frequently, particular replacements may direct new genes into novel functions. Subfunctionalisation can also occur if parental and duplicates retain function but become distinct and complementary in their spatiotemporal pattern of expression.

Mammalian MT1 and MT2 are conserved proteins that play a critical role in heavy-metal homeostasis and are transcriptionally induced by metal [51] and glucocorticoids [52]. While most of the mammals show a single MT1 and MT2 copy that evolved through a duplication event, primates harbor multiple MT1 copies (Fig. 1). In humans, MT1 expansion resulted in a total of 13 tandemly arranged genes, five of which are known or predicted to be pseudogenes, while the remaining eight are still functionally active (Fig. 2). The pseudogenisation process has occurred by the accumulation of loss-of-function mutations mainly by replacing critical metal-binding cysteines or incorporated aromatic amino acids in the protein sequence (Fig. 5). The most recently documented of these pseudogenes, MT1L [53], shows a unique mutation in which an invariant cysteine is replaced by a premature stop codon (Cys26Stop) resulting in a truncated protein.

Since its discovery [54], MT3 has been frequently associated with the protection against neuronal injury [15,16]. The mammalian MT3 protein shows a characteristic insertion of six residues at the a-domain when compared to that of MT1 and MT2 (Fig. 3) and an extra residue in the b domain (Thr), which is responsible for neuron growth inhibitory activity in Alzheimer disease [35,36]. Although the expression of MT3 has been almost exclusively related to brain tissues, we demonstrate that MT3 is a ubiquitously expressed gene. These results would drive future investigations on the involvement of MT3 in other cellular processes. In this regard, it is worth mentioning that Mt3 associates with other proteins in mouse brains as part of a multiprotein complex [55] suggesting function diversification and involvement in various physiological processes.

The most recently discovered family member, MT4, retains a high degree of conservation between humans and mice (Fig. 4), yet it shows the highest sequence divergence when compared with any other MT family member. However, the detection of structurally disrupting mutations at polymorphic proportions (Table 1) predicts that MT4 is inactive in some individuals. If that is the case, might the role of MT4 be performed by another family member? To address this possibility, the metal binding properties of MT4 were

explored in the literature data. It has been shown that mouse Mt4 retains the capacity to bind Zn [17,37], Cd and Cu as the ubiquitously expressed Mt1/Mt2, although the affinity to Cu is higher [37,56]. Furthermore, it showed similar characteristic metal-thiolate clusters and solvent accessibility as Mt1 [57]. Extrapolating these properties to the human protein, for which no data of such detail are available, it is tempting to assume that the loss (pseudogenisation) of MT4 can be compensated by functional equivalents. In this context, MT1 and MT2 would be the most likely candidates for a number of reasons. First, the metal binding properties of Mt1 and Mt2 overlap that of Mt4 in mice [57]. Second, it has been demonstrated that some MT1 duplicates have cellular specificity [44,46] and some of them are expressed in epithelium. Third, previous experiments in *Drosophila melanogaster* demonstrated that the number of functional gene duplications correlates to the resistance to Cd [13,58] and Cu [13] as a direct consequence of the increased gene expression. Taken these data together, is thus possible that additional MT1 duplicates may have assumed the specialized role of the inactivated protein in humans. In such a case, it is likely that some MT1 genes may act as compensatory backup copies that replace MT4 in individuals carrying deleterious mutations. Accordingly, it is thus further possible to infer that the compensatory mechanism implies the regulation of gene expression as observed in *D. melanogaster*. Tracing the evolutionary history of a gene after duplication often leaves more questions than answers. Some of those answers are easily obtained by direct read of DNA or protein sequences while some other depend on additional information which was here gathered for the MT family. Data on phylogeny, expression and intra-specific polymorphic information are necessary to drive hypotheses regarding the gain, loss and change of function of duplicated genes. The application of such a network of information, as is the case of this study, is thus necessary to extend the knowledge about the evolutionary fate of genes originated by duplication events.

Author Contributions

Conceived and designed the experiments: AA LA. Performed the experiments: AM JC RM LA. Analysed the data: AM JC RM RMS AA LA. Contributed reagents/materials/analysis tools: RM AA LA. Wrote the paper: AA LA.

Funding

This work was supported by Fundação para a Ciência e a Tecnologia grant (FCOMP-01-0124-FEDER-007167). LA, RM, and RS are supported by Fundação para a Ciência e a Tecnologia Ciência 2007 and by European Social Fund. Institute of Molecular Pathology and Immunology of the University of Porto is an Associate Laboratory of the Portuguese Ministry of Science, Technology and Higher Education and is partially supported by Fundacão para a Ciência e a Tecnologia. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Competing Interests: The authors have declared that no competing interests exist.

Figures and Tables

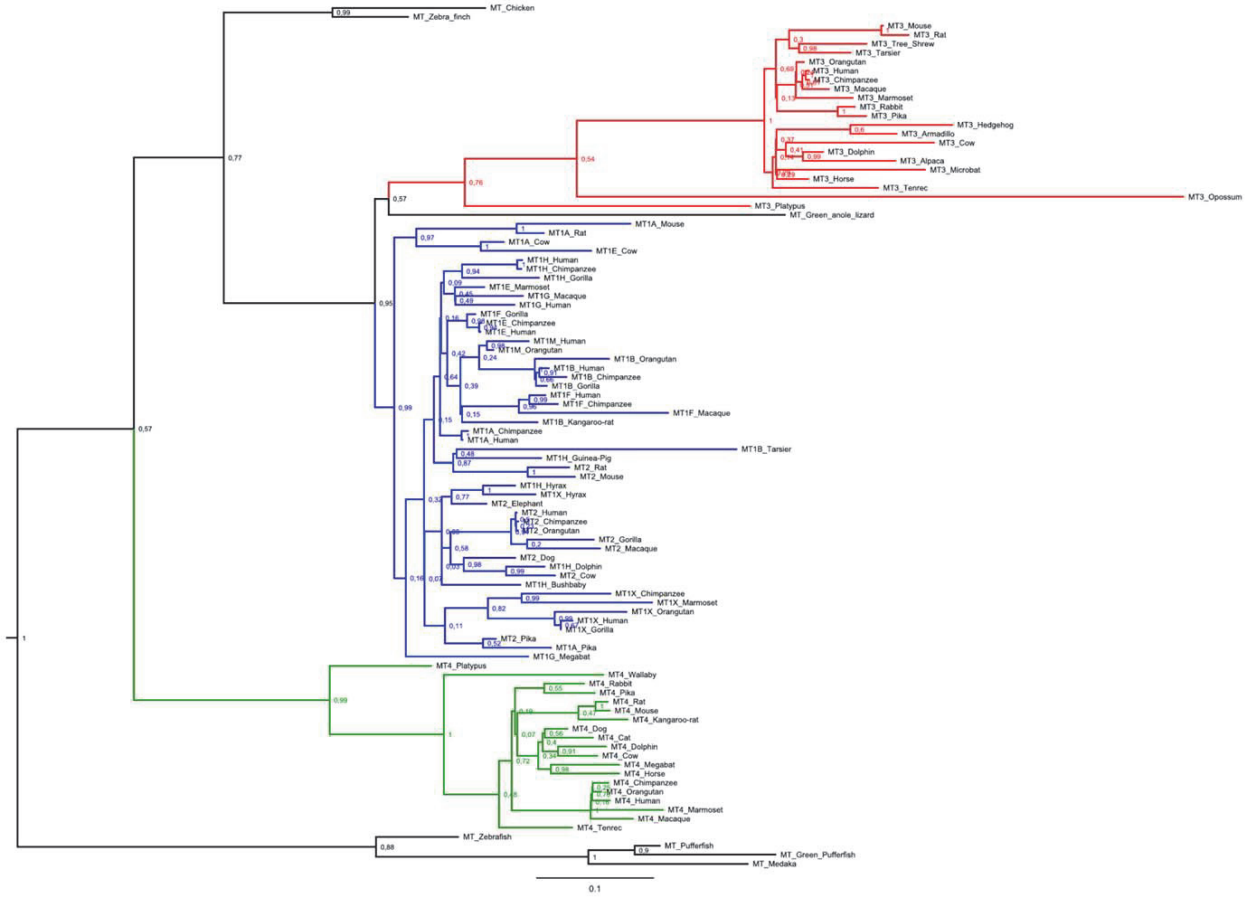


Figure 1: Bayesian phylogenetic analysis of the MT family. The gene tree was constructed using coding sequences from the Ensembl database (Table S1). MT1/MT2, MT3 and MT4 clusters are represented in blue, red and green, respectively. Posterior probability values are given for branch support. Scale bar stands for the number of replacements per site.

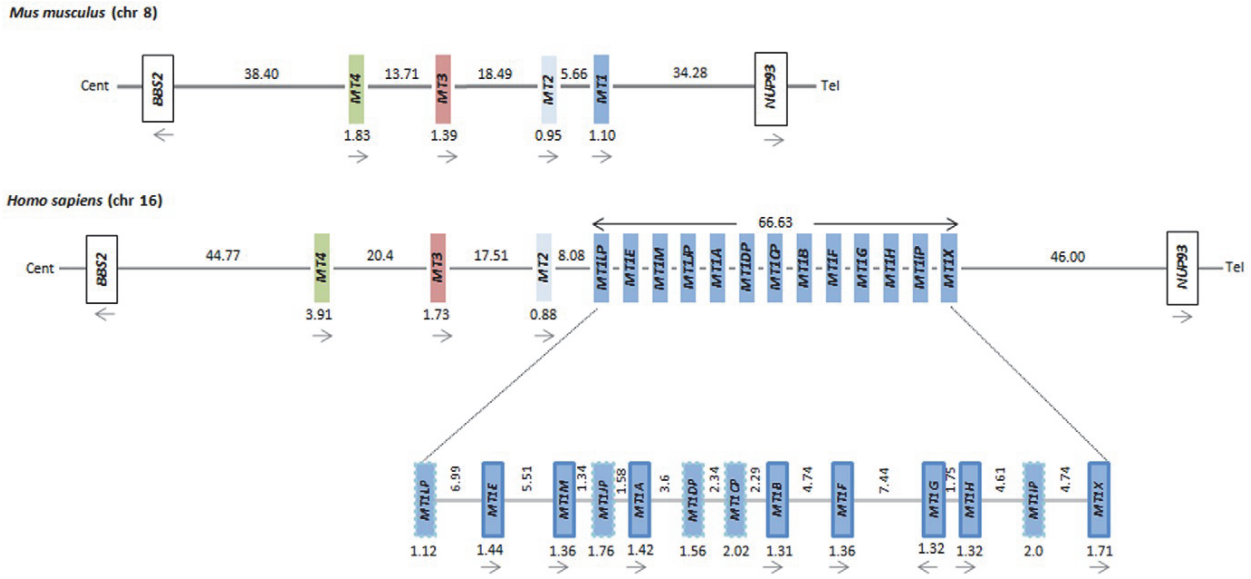


Figure 2: Illustration of the MT family in humans and mice. In mice, the family comprises four functionally active genes (mt1 to mt4). In humans, the family harbours a single-copy of MT2, MT3 and MT4, and by a tandemly duplicated array of the MT1 duplicates spanning about 66.6 Kb, where eight active genes (MT1A to MT1J, MT1L, MT1M and MT1X) and five pseudogenes (MT1L, MT1J, MT1D, MT1C and MT1I) are known. The direction of transcription for each active gene is indicated by an arrow. Genes are coloured as follows: MT1 (dark blue), MT2 (light blue), MT3 (red) and MT4 (green). Pseudogenes are represented by dashed boxes. Numbers corresponding to the sizes of gene and intergenic regions are given in Kb.

A

```
                                  β                                α
Hs_MT1E    MDPN-CSCATGGSCTCAGSCKCKECKCTSCKKSCCSCCPVGCAKCAQGCVCKG......ASEKCSCCA 61
Mm_Mt1     MDPN-CSCSTGGSCTCTSSCACKNCKCTSCKKSCCSCCPVGCSKCAQGCVCKG......AADKCTCCA 61
Hs_MT2     MDPN-CSCAAGDSCTCAGSCKCKECKCTSCKKSCCSCCPVGCAKCAQGCICKG......ASDKCSCCA 61
Mm_Mt2     MDPN-CSCASDGSCSCAGACKCKQCKCTSCKKSCCSCCPVGCAKCSQGCICKE......ASDKCSCCA 61
Hs_MT3     MDPETCPCPSGGSCTCADSCKCEGCKCTSCKKSCCSCCPAECEKCAKDCVCKGGEAAEAEAEKCSCCQ 68
Mm_Mt3     MDPETCPCPTGGSCTCSDKCKCKGCKCTNCKKSCCSCCPAGCEKCAKDCVCKGEEGAKAEAEKCSCCQ 68
Hs_MT4     MDPRECVCMSGGICMCGDNCKCTTCNCKTCRKSCCPCCPPGCAKCARGCICKG......GSDKCSCCP 62
Mm_Mt4     MDPGECTCMSGGICICGDNCKCTTCSCKTCRKSCCPCCPPGCAKCARGCICKG......GSDKCSCCP 62
```

B

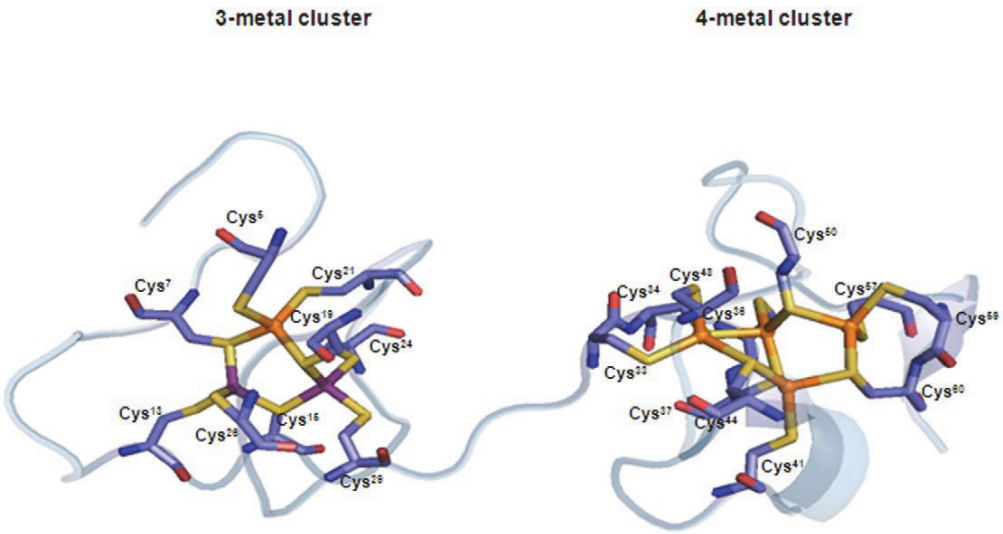


Figure 3: Sequence comparison and structural features of MT proteins. (A) Sequence alignment for human (Hs) and mouse (Mm) proteins. Residues spanning a and b domains are indicated as well as are the invariant metal-binding cysteine residues. Sequences were aligned with ClustalW [59] (B) Structural representation of the mouse Mt2 (PDB 4mt2) showing the two metal-binding clusters and the detailed spatial organisation of the cysteine residues with the S-atoms oriented towards metal ions. Elements are coloured as follows: S (yellow), O (red), N (dark blue), Zn (purple) and Cd (orange).

	MT4	MT3	MT2	MT1E	MT1M	MT1A	MT1B	MT1F	MT1G	MT1H	MT1X
Mt1	59	70	82	85	79	82	79	82	82	79	80
Mt2	64	66	86	85	74	82	79	80	84	84	85
Mt3	60	87	69	75	70	72	70	72	70	70	67
Mt4	94	61	66	64	59	64	60	60	64	66	66

50-60
61-70
71-80
81-90
91-100

Figure 4: The amino acid identity matrix for human and mouse proteins. Amino acid identity for each pairwise comparison between a human and a mouse protein is given as percentage values.

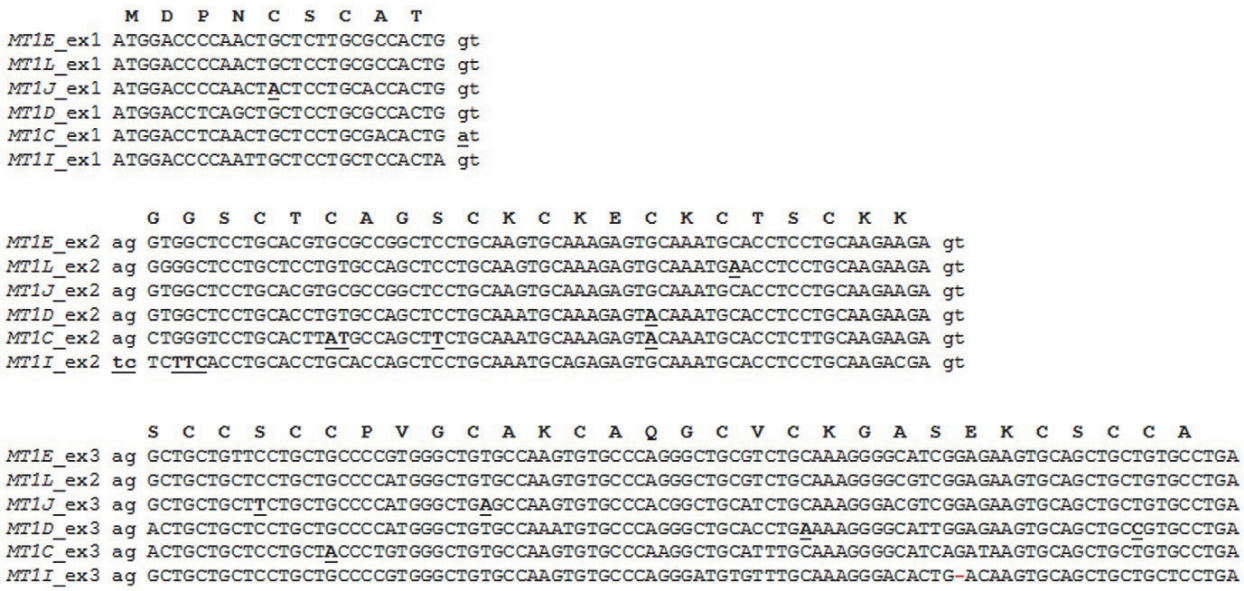


Figure 5: The reconstructed open reading frame of MT1 pseudogenes. The reconstruction was performed using the genomic sequence of human MT1 pseudogenes to a homology-based comparison with the functional MT1E. Each exon is shown in a separate row. Amino-acids are represented in single-letter code above the corresponding codon. Strong deleterious mutation candidates (loss of invariant cysteines, gain of aromatic residues, indels and splice-site mutations) are underlined.

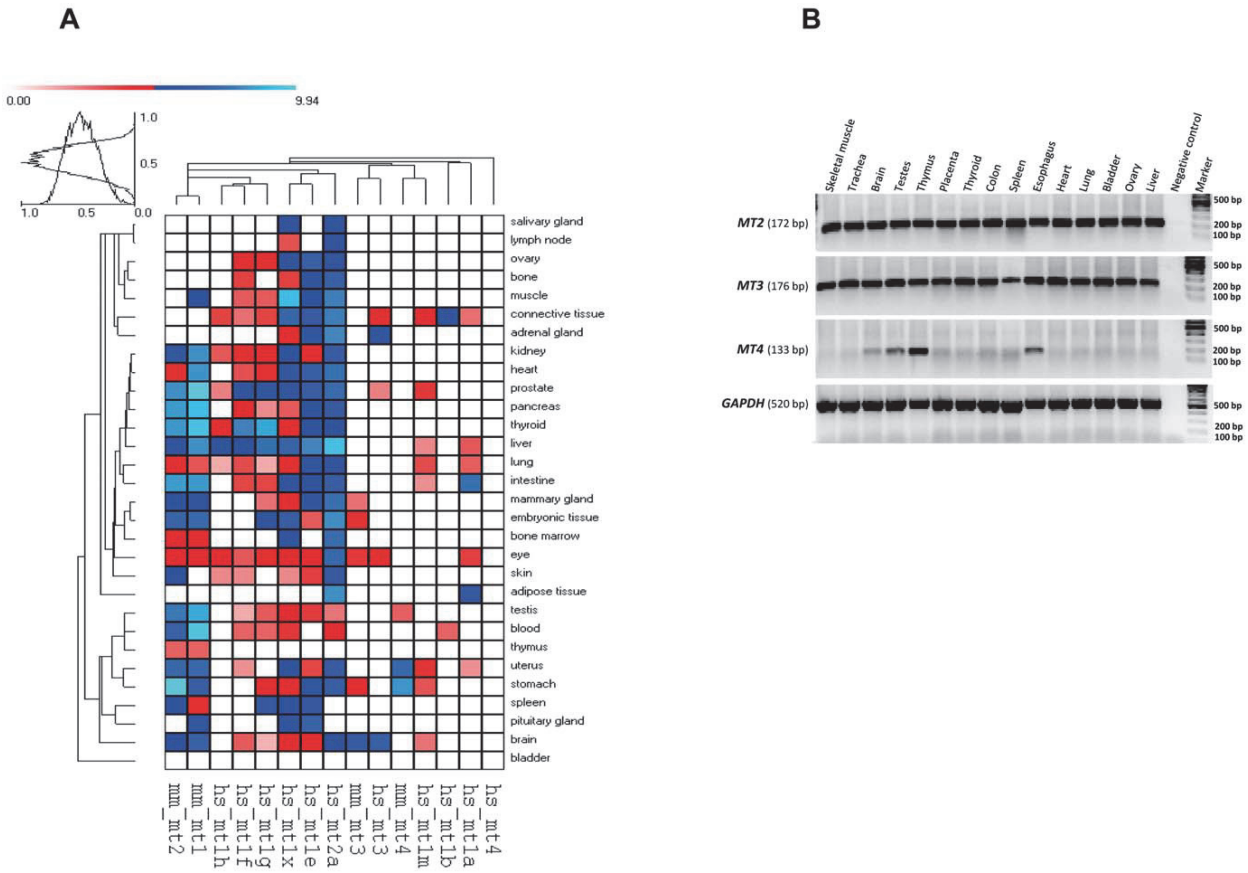


Figure 6: Expression profile of MT genes. (A) Expression of the MT genes in human and mouse tissues represented as $\log_2$ of EST counts by colour coding. (B) RT-PCR analysis of human MT2, MT3 and MT4 in 15 tissues using GAPDH as control.

Table 1. Nonsynonymous replacements at the human MT genes annotated in Ensembl database.

Gene	Nonsynonymous variants
MT1E	Asn40Ser (rs12051120) Arg46Lys (rs34166523)
MT1M	Thr20Lys (rs1827210)
MT1A	Thr27Asn (rs11640851) Lys51Arg (rs8052394)
MT1B	Cys19Ser (rs61744104)
MT1F	-
MT1G	-
MT1H	Gly17Arg (rs9934181) MT1X
MT4	Cys30Tyr (rs666636) Arg31Trp (rs666647) Gly48Asp (rs11643815)

Table 2 : Hapmap allelic frequencies of MT4 Tyr30 (rs666636) and Trp31 (rs666647) in human population.

Population	TYR30-A allele	TRP31-T allele
African		
ASW	0.189	0.189
LWK	0.303	0.244
MKK	0.155	0.15
YRI	0.161	0.124
Asian		
CHB	0.298	0.298
CHD	0.288	0.282
JPT	0.267	0.265
GIH	0.045	0.045
European		
CEU	0.085	0.058
TSI	0.034	0.034
American		
MEX	0.05	0.05

Population description as indicated in Hapmap: ASW, African ancestry in Southwest USA; CEU, Utah residents with Northern and Western European ancestry from the CEPH collection; CHB, Han Chinese in Beijing, China; CHD, Chinese in Metropolitan Denver, Colorado; GIH, Gujarati Indians in Houston, Texas; JPT, Japanese in Tokyo, Japan; LWK, Luhya in Webuye, Kenya; MEX, Mexican ancestry in Los Angeles, California; MKK, Maasai in Kinyawa, Kenya; TSI, Toscans in Italy; and YRI, Yoruba in Ibadan, Nigeria.

Supporting Information

Figure S1: Maximum likelihood analysis of the MT family. The gene tree was constructed using coding sequences from the Ensembl database (Table S1). MT1/MT2, MT3 and MT4 clusters are represented in blue, red and green, respectively. Branch support was estimated by bootstrap. Scale bar: number of replacements per site.

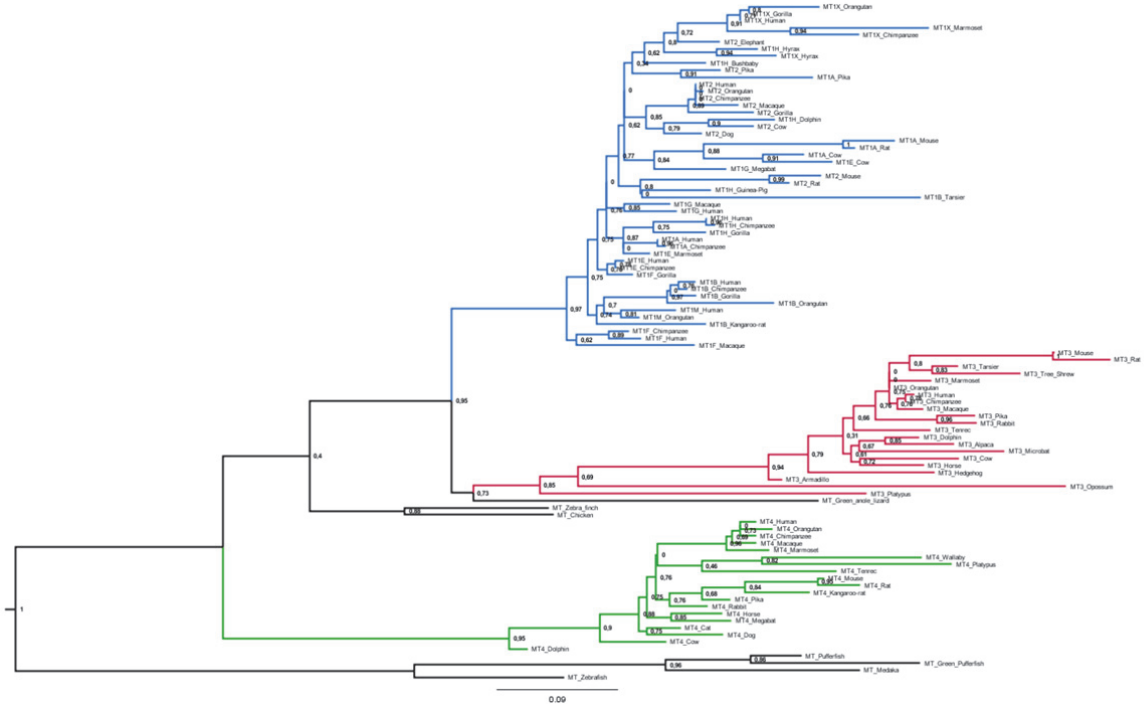


Table S1: Metallothionein gene transcripts annotated in the Ensembl database (release 56, Sep 2009) as human orthologues of MT1E, MT1M, MT1A, MT1B, MT1F, MT1G, MT1H, MT1X, MT2, MT3, and MT4 in mammals. MT sequences from fishes, birds and reptiles are shown at the bottom of the table.

GENE	SCIENTIFIC NAME	COMMON NAME	ENSEMBL TRANSCRIPT
MT1E	*Bos taurus*	Cow	ENSBTAT00000046456
	Pan troglodytes	Chimpanzee	ENSPTRT00000063817
	Homo sapiens	Human	ENST00000306061
	Callithrix jacchus	Marmoset	ENSCJAT00000024529
MT1M	*Homo sapiens*	Human	ENST00000379818
	Pongo pygmaeus	Orangutan	ENSPPYT00000029425
MT1A	*Bos taurus*	Cow	ENSBTAT00000002092
	Ochotona princeps	Pika	ENSOPRT00000006730
	Pan troglodytes	Chimpanzee	ENSPTRT00000015005
	Homo sapiens	Human	ENST00000290705
	Mus musculus	Mouse	ENSMUST00000034215
	Rattus norvegicus	Rat	ENSRNOT00000038212
MT1B	*Pan troglodytes*	Chimpanzee	ENSPTRT00000015003
	Gorilla gorilla	Gorilla	ENSGGOT00000009888
	Homo sapiens	Human	ENST00000334346

	Pongo pygmaeus	Orangutan	ENSPPYT00000009031
	Tarsius syrichta	Tarsier	ENSTSYT00000007085
	Dipodomys ordii	Kangaroo rat	ENSDORT00000003898
MT1F	*Pan troglodytes*	Chimpanzee	ENSPTRT00000015004
	Gorilla gorilla	Gorilla	ENSGGOT00000009872
	Homo sapiens	Human	ENST00000334350
	Macaca mulatta	Macaque	ENSMMUT00000039321
MT1G	*Pteropus vampyrus*	Megabat	ENSPVAT00000016198
	Homo sapiens	Human	ENST00000379811
	Macaca mulatta	Macaque	ENSMMUT00000022653
MT1H	*Tursiops truncatus*	Dolphin	ENSTTRT00000013550
	Procavia capensis	Hyrax	ENSPCAT00000013294
	Otolemur garnettii	Bushbaby	ENSOGAT00000008589
	Pan troglodytes	Chimpanzee	ENSPTRT00000014999
	Gorilla gorilla	Gorilla	ENSGGOT00000016717
	Homo sapiens	Human	ENST00000332374
	Cavia porcellus	Guinea Pig	ENSCPOT00000011224
MT1X	*Procavia capensis*	Hyrax	ENSPCAT00000007828
	Pan troglodytes	Chimpanzee	ENSPTRT00000067756
	Gorilla gorilla	Gorilla	ENSGGOT00000015825
	Homo sapiens	Human	ENST00000394485
	Callithrix jacchus	Marmoset	ENSCJAT00000023761
	Pongo pygmaeus	Orangutan	ENSPPYT00000002180
MT2	*Bos taurus*	Cow	ENSBTAT00000034373
	Canis familiaris	Dog	ENSCAFT00000014487
	Ochotona princeps	Pika	ENSOPRT00000001091
	Pan troglodytes	Chimpanzee	ENSPTRT00000014994
	Gorilla gorilla	Gorilla	ENSGGOT00000012065
	Homo sapiens	Human	ENST00000245185
	Macaca mulatta	Macaque	ENSMMUT00000031285
	Pongo pygmaeus	Orangutan	ENSPPYT00000008667
	Loxodonta africana	Elephant	ENSLAFT00000012833
	Mus musculus	Mouse	ENSMUST00000034214
	Rattus norvegicus	Rat	ENSRNOT00000067391
MT3	*Vicugna pacos*	Alpaca	ENSVPAT00000003704
	Bos taurus	Cow	ENSBTAT00000022460
	Echinops telfairi	Lesser hedgehog tenrec	ENSETET00000000091
	Tursiops truncatus	Dolphin	ENSTTRG00000013549
	Myotis lucifugus	Microbat	ENSMLUT00000005003
	Dasypus novemcinctus	Armadillo	ENSDNOT00000015796
	Erinaceus europaeus	Hedgehog	ENSEEUT00000003639
	Ochotona princeps	Pika	ENSOPRT00000006724
	Oryctolagus cuniculus	Rabbit	ENSOCUT00000016238
	Equus caballus	Horse	ENSECAT00000015904
	Pan troglodytes	Chimpanzee	ENSPTRT00000014993
	Homo sapiens	Human	ENST00000200691
	Macaca mulatta	Macaque	ENSMMUT00000010829

	Callithrix jacchus	*Marmoset*	ENSCJAT00000024548
	Pongo pygmaeus	Orangutan	ENSPPYT00000008666
	Tarsius syrichta	Tarsier	ENSTSYT00000001126
	Mus musculus	Mouse	ENSMUST00000034211
	Rattus norvegicus	Rat	ENSRNOT00000025669
	Tupaia belangeri	Tree Shrew	ENSTBET00000016034
	Monodelphis domestica	Opossum	ENSMODT00000040198
	Ornithorhynchus anatinus	Platypus	ENSOANT00000029863
MT4	*Bos taurus*	Cow	ENSBTAT00000020072
	Echinops telfairi	Lesser hedgehog tenrec	ENSETET00000005977
	Felis catus	Cat	ENSFCAT00000008674
	Canis familiaris	Dog	ENSCAFT00000014504
	Tursiops truncatus	Dolphin	ENSTTRT00000013548
	Pteropus vampyrus	Megabat	ENSPVAT00000014588
	Macropus eugenii	Wallaby	ENSMEUT00000011205
	Ochotona princeps	Pika	ENSOPRT00000006718
	Oryctolagus cuniculus	Rabbit	ENSOCUT00000017267
	Equus caballus	Horse	ENSECAT00000014386
	Pan troglodytes	Chimpanzee	ENSPTRT00000014992
	Homo sapiens	Human	ENST00000219162
	Macaca mulatta	Macaque	ENSMMUT00000022652
	Callithrix jacchus	*Marmoset*	ENSCJAT00000024551
	Pongo pygmaeus	Orangutan	ENSPPYT00000008665
	Dipodomys ordii	Kangaroo rat	ENSDORT00000015428
	Mus musculus	Mouse	ENSMUST00000034207
	Rattus norvegicus	Rat	ENSRNOT00000025694
	Ornithorhynchus anatinus	Platypus	ENSOANT00000029863

Fishes

Danio rerio	Zebrafish	ENSDART00000061007
Oryzias latipes	Medaka	ENSORLT00000019509
Tetraodon nigroviridis	Green putterfish	ENSTNIT00000011862
Takifugu rubripes	Putterfish	ENSTRUT00000022487

Birds

Taeniopygia guttata	Zebra finch	ENSTGUT00000006787
Gallus gallus	Chicken	ENSGALT00000023565

Reptiles

Anolis carolinensis	Anole lizard	ENSACAT00000007496

References

1. Hughes AL (2002) Adaptive evolution after gene duplication. Trends Genet 18:433–434.
2. Lynch M, Katju V (2004) The altered evolutionary trajectories of gene duplicates. Trends Genet 20: 544–549.
3. Lynch M (2007) The origins of genome architecture. Sunderland: Sinauer Associates. pp 389.
4. Ohno S (1970) Evolution by gene duplication. Heidelberg: Springer-Verlag.160 p.
5. Ohta T (1990) How gene families evolve. Theor Popul Biol 37: 213–219.
6. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155.
7. Ohta T (1987) Simulating evolution by gene duplication. Genetics 115:207–213.
8. Prince VE, Pickett FB (2002) Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet 3: 827–837.
9. Force A, Lynch M, Pickett FB, Amores A, Yan YL, et al. (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545.
10. Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. Genetics 154: 459–473.
11. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, et al. (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421: 63–66.
12. Ihmels J, Collins SR, Schuldiner M, Krogan NJ, Weissman JS (2007) Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss. Mol Syst Biol 3: 86.
13. Maroni G, Wise J, Young JE, Otto E (1987) Metallothionein gene duplications and metal tolerance in natural populations of Drosophila melanogaster. Genetics
117: 739–744.
14. Egli D, Domenech J, Selvaraj A, Balamurugan K, Hua H, et al. (2006) The four members of the Drosophila metallothionein family exhibit distinct yet overlapping roles in heavy metal homeostasis and detoxification. Genes to Cells 11: 647–658.
15. Chung RS, West AK (2004) A role for extracellular metallothioneins in CNS injury and repair. Neuroscience 123: 595–599.
16. Hozumi I, Inuzuka T, Hiraiwa M, Uchida Y, Anezaki T, et al. (1995) Changes of growth inhibitory factor after stab wounds in rat brain. Brain Res 688:143–148.
17. Quaife CJ, Findley SD, Erickson JC, Froelick GJ, Kelly EJ, et al. (1994) Induction of a New Metallothionein Isoform (MT-IV) Occurs during Differentiation of Stratified Squamous Epithelia. Biochemistry 33: 7250–7259.
18. Villarreal L, Tío L, Capdevila M, Atrian S (2006) Comparative metal binding and genomic analysis of the avian (chicken) and mammalian metallothionein. FEBS Journal 273: 523–535.
19. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, et al. (2009) Ensembl 2009.
Nucleic Acids Res 37: D690–697.
20. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.
21. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
22. Posada D (2008) jModelTest: phylogenetic model averaging. Mol Biol Evol 25:1253–1256.
23. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7: 214.
24. Rambaut A, Drummond AJ Tracer v1.4 [http://beast.bio.ed.ac.uk/Tracer].

25. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.

26. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2007) Database resources of the National Center for Biotechnology Information. Nucl Acids Res 35: D5–12.

27. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, et al. (2004) EnsMart: a generic system for fast and flexible access to biological data. Genome Res 14: 160–169.

28. Braun W, Vasak M, Robbins AH, Stout CD, Wagner G, et al. (1992) Comparison of the NMR solution structure and the x-ray crystal structure of rat metallothionein-2. Proc Natl Acad Sci U S A 89: 10124–10128.

29. Sali A, Blundell TL (1993) Comparative Protein Modelling by Satisfaction of Spatial Restraints. Journal of Molecular Biology 234: 779–815.

30. Azevedo L, Carneiro J, van Asch B, Moleirinho A, Pereira F, et al. (2009) Epistatic interactions modulate the evolution of mammalian mitochondrial respiratory complex components. BMC Genomics 10: 266.

31. Laukens D, Waeytens A, De Bleser P, Cuvelier C, De Vos M (2009) Human metallothionein expression under normal and pathological conditions: mecha- nisms of gene regulation based on *in silico* promoter analysis. Crit Rev Eukaryot Gene Expr 19: 301–317.

32. Waalkes MP, Goering PL (1990) Metallothionein and other cadmium-binding proteins: recent developments. Chem Res in Toxicol 3: 281–288.

33. Boulanger Y, Goodman CM, Forte CP, Fesik SW, Armitage IM (1983) Model for mammalian metallothionein structure. Proc Natl Acad Sci U S A 80:
1501–1505.

34. Duncan KER, Ngu TT, Chan J, Salgado MT, Merrifield ME, et al. (2006) Peptide Folding, Metal-Binding Mechanisms, and Binding Site Structures in
Metallothioneins. Exp Biol Med 231: 1488–1499.

35. Cai B, Zheng Q, Teng X-C, Chen D, Wang Y, et al. (2006) The role of Thr5 in human neuron growth inhibitory factor. J Biol Inorg Chem 11: 476–482.

36. Romero-Isart N, Jensen LT, Zerbe O, Winge DR, Vasak M (2002) Engineering
of Metallothionein-3 Neuroinhibitory Activity into the Inactive Isoform Metallothionein-1. J Biol Chem 277: 37023–37028.

37. Tio L, Villarreal L, Atrian S, Capdevila M (2004) Functional differentiation in the mammalian metallothionein gene family: metal binding features of mouse MT4 and comparison with its paralog MT1. J Biol Chem 279: 24403–24413.

38. Greer JM, Puetz J, Thomas KR, Capecchi MR (2000) Maintenance of functional equivalence during paralogous Hox gene evolution. Nature 403:
661–665.

39. Gu Z, Rifkin SA, White KP, Li W-H (2004) Duplicate genes increase gene expression diversity within and between species. Nat Genet 36: 577–579.

40. Hamilton AT, Huntley S, Tran-Gyamfi M, Baggott DM, Gordon L, et al. (2006) Evolutionary expansion and divergence in the ZNF91 subfamily of primate-specific zinc finger genes. Genome Res 16: 584–594.

41. Byrne KP, Wolfe KH (2007) Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. Genetics 175: 1341–1350.

42. Zou S, Kamei H, Modi Z, Duan C (2009) Zebrafish IGF genes: gene
duplication, conservation and divergence, and novel roles in midline and notochord development. PLoS One 4: e7026.

43. Masters BA, Quaife CJ, Erickson JC, Kelly EJ, Froelick GJ, et al. (1994) Metallothionein III is expressed in neurons that sequester zinc in synaptic vesicles. J Neurosci 14: 5844–5857.

44. Schmidt CJ, Hamer DH (1986) Cell specificity and an effect of ras on human metallothionein gene

expression.Proc Natl Acad Sci U S A 83: 3346–3350.

45. Stennard (1994) Biochim Biophys Acta.
46. Cherian MG, Jayasurya A, Bay BH (2003) Metallothioneins in human tumors and potential roles in carcinogenesis. Mutat Res 533: 201–209.
47. Haq F, Mahoney M, Koropatnick J (2003) Signaling events for metallothionein induction. Mutat Res 533: 211–226.
48. Vasak M, Hasler DW (2000) Metallothioneins: new functional and structural insights. Curr Opin Chem Biol 4: 177–183.
49. Karin M, Eddy RL, Henry WM, Haley LL, Byers MG, et al. (1984) Human metallothionein genes are clustered on chromosome 16. Proc Natl Acad Sci U S A

81: 5494–5498.

50. Rodewald HR, Paul S, Haller C, Bluethmann H, Blum C (2001) Thymus medulla consisting of epithelial islets each derived from a single progenitor. Nature 414: 763–768.
51. Beach LR, Palmiter RD (1981) Amplification of the metallothionein-I gene in cadmium-resistant mouse cells. Proc Natl Acad Sci U S A 78: 2110–2114.
52. Hager LJ, Palmiter RD (1981) Transcriptional regulation of mouse liver metallothionein-I gene by glucocorticoids. Nature 291: 340–342.
53. Hahn Y, Lee B (2006) Human-specific nonsense mutations identified by genome sequence comparisons. Human Genetics 119: 169–178.
54. Palmiter RD, Findley SD, Whitmore TE, Durnam DM (1992) MT-III, a brain- specific member of the metallothionein gene family. Proceedings of the National Academy of Sciences of the United States of America 89: 6333–6337.
55. Lahti DW, Hoekman JD, Tokheim AM, Martin BL, Armitage IM (2005) Identification of mouse brain proteins associated with isoform 3 of metallothio- nein. Protein Sci 14: 1151–1157.
56. Meloni G, Zovo K, Kazantseva J, Palumaa P, Vasak M (2006) Organization and assembly of metal-thiolate clusters in epithelium-specific metallothionein-4. J Biol Chem 281: 14588–14595.
57. Cai B, Zheng Q, Huang Z-X (2005) The properties of the metal–thiolate clusters in recombinant mouse metallothionein-4. Protein J 24: 327–336.
58. Otto E, Young JE, Maroni G (1986) Structure and expression of a tandem duplication of the Drosophila metallothionein gene. Proc Natl Acad Sci U S A 83: 6025–6029.
59. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res 22: 4673–4680.

8. Publication II: The Evolutionary Portrait of Metazoan NAD Salvage

The Evolutionary Portrait of Metazoan NAD Salvage

João Carneiro[1,2#], Sara Duarte-Pereira[1#], Luísa Azevedo[1], L. Filipe C. Castro[3], Paulo Aguiar[4], Irina S. Moreira[5], António Amorim[1,2] and Raquel M. Silva[1*]

[1]IPATIMUP-Institute of Molecular Pathology and Immunology of the University of Porto, Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

[2]Faculty of Sciences, University of Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal

[3]Interdisciplinary Centre for Marine and Environmental Research (CIIMAR), CIMAR Associate Laboratory, University of Porto, Portugal

[4]CMUP - Centro de Matemática da Universidade do Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal

[5]REQUIMTE - Rede de Química e Tecnologia, Faculty of Sciences, University of Porto, Rua do Campo Alegre 687, 4169-007 Porto, Portugal

[#]These authors contributed equally to this work and are listed by alphabetical order.

*Corresponding author:
Raquel M. Silva
IPATIMUP - Institute of Molecular Pathology and Immunology of the University of Porto
Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal
Tel: +351 225570700
Fax: +351 225570799
Email: raquelsilva@ua.pt

Keywords: NAD salvage, NAMPT, nicotinamidase, evolution, homology modeling, protein-ligand docking, metabolism

Citation: Carneiro J, Duarte-Pereira S, Azevedo L, Castro LFC, Aguiar P, et al. (2013) The Evolutionary Portrait of Metazoan NAD Salvage. PLoS ONE 8(5): e64674. doi:10.1371/journal.pone.0064674
Editor: Andrew R. Dalby, University of Westminster, United Kingdom
Received: February 3, 2013; Accepted: April 16, 2013; Published: May 28, 2013

Abstract

Nicotinamide Adenine Dinucleotide (NAD) levels are essential for cellular homeostasis and survival. Main sources of intracellular NAD are the salvage pathways from nicotinamide, where Nicotinamide phosphoribosyltransferases (NAMPTs) and Nicotinamidases (PNCs) have a key role. NAMPTs and PNCs are important in aging, infection and disease conditions such as diabetes and cancer. These enzymes have been considered redundant since either one or the other exists in each individual genome. The co-occurrence of *NAMPT* and *PNC* was only recently detected in invertebrates though no structural or functional characterization exists for them. Here, using expression and evolutionary analysis combined with homology modeling and protein-ligand docking, we show that both genes are expressed simultaneously in key species of major invertebrate branches and emphasize sequence and structural conservation patterns in metazoan NAMPT and PNC homologues. The results anticipate that NAMPTs and PNCs are simultaneously active, raising the possibility that NAD salvage pathways are not redundant as both are maintained to fulfil the requirement for NAD production in some species.

Introduction

Nicotinamide Adenine Dinucleotide (NAD) is an essential molecule to cells. As a cofactor in redox reactions, NAD regulates the metabolism and energy production and, as a substrate for NAD-consuming enzymes such as poly(ADP-ribose) polymerases (PARPs) and sirtuins, NAD is involved in DNA repair, transcriptional silencing and cell survival [1]. To maintain adequate NAD levels, several routes are used for NAD synthesis that depend on distinct precursors: *de novo* pathways synthesize NAD from tryptophan or aspartic acid whereas salvage pathways recycle NAD from nicotinamide (Nam), nicotinic acid (Na) and their ribosides [2-4].

The nicotinamide salvage pathway is the major source of intracellular NAD in humans [5,6] and is also required for growth in several microorganisms [7-10]. NAD salvage from Nam is a two- or four-step reaction, in which the rate-limiting enzymes and functional homologues are, respectively, nicotinamide phosphoribosyltransferases (NAMPTs) and nicotinamidases (PNCs) [11-13]. In humans, *NAMPT* is widely studied due to its involvement in inflammation and disease such as cancer [14,15]. In contrast, humans lack nicotinamidase but expression of the *Drosophila* Pnc protects human neuronal cells from death originated by oxidative stress [16]. Moreover, an increased Pnc1 and sirtuin activity confers protection to proteotoxic stress in yeast and *C. elegans* [17,18]. The yeast Pnc1 is a biomarker of stress and a regulator of sirtuin activity [11,18], and thus, most studies in yeast and invertebrates have focused in the link between these enzymes and aging [16,19]. Notwithstanding, despite their importance to major cellular processes, there is a poor functional characterization of nicotinamidases [20,21] and their role in infection has been less explored [7,8,22].

NAMPTs and PNCs act as regulators of enzymes from similar pathways, coordinating the overall metabolism and stress responses [23]. Moreover, both are pharmacologically relevant. NAMPT inhibitors are being used in clinical trials as anti-cancer agents [24-27] and nicotinamidases are attractive targets to the development of drugs for infectious diseases and anti-parasitic therapies [7,8,22,28-30].

NAMPTs and PNCs do not co-occur in all organisms and, until very recently, lineages with both *NAMPT* and *PNC* had been only found in bacteria and algae [30-32]. *NAMPT* was thought to be absent from invertebrates but the discovery that *NAMPT* homologues are present in several invertebrate species and that some species have both *NAMPT* and *PNC* homologues [33] challenged the classical view that these enzymes are redundant and mutually exclusive [1], emphasizing the need for studies characterizing the structural and functional properties of these enzymes.

Motivated by the lack of information for *NAMPT* and *PNC* homologues in relevant invertebrate species, which would render the biological meaning of simultaneous *versus* unique occurrence of these proteins more evident, we carried out an integrated study to

establish gene expression, amino acid conservation and structural comparisons. We provide experimental evidence that both genes are expressed simultaneously in key invertebrate species. In addition, evolutionary conserved patterns at the amino acid sequence and at the structural levels were detected. Also, using homology modeling and protein-ligand docking, we identify the amino acids that bind Nam in the active sites of invertebrate NAMPTs and PNCs. Taken together, the results suggest that invertebrate NAMPTs and PNCs are concurrently functional and, thus, that both NAD salvage pathways might not be redundant.

Results

Expression of invertebrate NAMPTs and PNCs

NAMPT homologues have been previously found in the vibriophage KVP40 [34], bacteria [10,32], and the unicellular green algae *Chlamydomonas reinhardtii* [31], motivating the search for *NAMPT* homologues in invertebrates, some of which simultaneously have *PNC* sequences [33] (Table S1). No recognizable *NAMPT* homologue has been detected so far in representative species of the phyla Arthropoda or Nematoda, although *NAMPT* and *PNC* were found in more basal lineages such as the choanoflagellate *Monosiga brevicollis* and the sea anemone *N. vectensis* (Figure 1). Such phylogenetic distribution is consistent with a scenario where both genes were present in the Metazoa ancestor and were selectively lost in specific lineages, as evidenced by the different patterns in protostomes. Namely, both genes were found in lophotrochozoans that includes mollusks (*Lottia gigantea*) and annelids (*C. teleta* and *Helobdella robusta*), and the absence of *NAMPT* was observed in ecdysozoans such as nematodes and arthropods. In deuterostomes, which comprises chordates, hemichordates and echinoderms, both genes were likely present in early lineages, which is supported by the evidence from the extant *B. floridae*, *Saccoglossus kowaleskii* and *S. purpuratus* species, but *NAMPT* was secondarily lost in the urochordate *Ciona intestinalis* while *PNC* was lost in vertebrates (Figure S1). RT-PCR of selected species showed that both *NAMPT* and *PNC* genes are expressed in the adult forms of *Branchiostoma floridae* (Cephalochordata)*, Strongylocentrotus purpuratus* (Echinodermata), *Capitella teleta* (Annelida) and *Nematostella vectensis* (Cnidaria) (Figure 1). In addition, available EST (Expressed Sequence Tag) data indicates that *NAMPT* and *PNC* genes are also co-expressed during developmental stages (Table S2), suggesting a widespread usage of both Nam salvage pathways across Metazoa.

Evolutionary divergence of NAMPTs and PNCs

We have further characterized the evolutionary divergence of NAMPT and PNC homologues, measured as protein distances calculated from amino acid sequence alignments (Figure 2). The resulting matrix (Figure 2) showed that NAMPT is conserved, even when large evolutionary distances are considered. For example, the divergence between the human and cnidarian (*N. vectensis*) NAMPT homologues is about 50%, as much as when compared with amphioxus (*B. floridae*). Among invertebrates the sequences showing the smallest divergence are from *N. vectensis* and *C. teleta* (31.2%). Conversely, PNC sequences are highly divergent even in closely related species, as shown for the annelids *C. teleta* and *H. robusta*, or the basal chordates *B. floridae* and *C. intestinalis*. Curiously, the smallest divergence between PNC sequences was found for *C. teleta* and *B. floridae* (51.3%). This trend was also evident when we plotted protein distances taking implicitly in consideration the evolutionary divergence time between each pair of species studied (Figure S2 and Table S3). Analyses of protein distances (*pd*) indicated that NAMPT homologues are considerably more conserved (*pd* = 0.447±0.116) than PNC (*pd* = 0.842±0.151) (mean±std), which is remarkable for species spanning over 1000 million years of divergence (Table S3). For PNC proteins, in addition to the larger values, no correlation with evolutionary distance was observed, while NAMPT distances were smaller and increased consistently with the evolutionary distance (*ed*) between species. The Kendall rank correlation coefficient was used to measure the dependence between *pd* and *ed*, showing no relevant dependence between both quantities for PNC (τ=–0.052). However, for NAMPT both quantities vary consistently (τ=0.413).

Motif conservation in NAMPTs and PNCs

We next used the previously constructed amino acid sequence alignments dataset to search for conserved motifs in NAMPT and PNC homologues. In line with the aforementioned results, analyses of NAMPT sequences (Figure 3A) revealed conserved amino acid motifs surrounding catalytic residues [24,25,35-37] Tyr18, Phe193, Asp219, His247, Asp279, Asp313, corresponding to the boxed amino acids in Figure 3. As well, Asp16 and Arg311, Gly353 and Asp354, and Gly384 that bind nicotinamide, ribose or phosphate, respectively, are preserved and the additional NMN interacting residues Arg196 and Gly383 in rat NAMPT [25] are present in all sequences analysed. The amino acid stretches that represent the dimer interface are also conserved in invertebrate NAMPTs (Figure 3A and Figure S3), as previously shown for vertebrates [25].

Similar analyses on PNC homologues showed that, while overall amino acid sequence identity is low (Figure 3B), motifs surrounding metal-binding and catalytic residues (boxed amino acids) show up. Indeed, all PNC sequences have conserved residues that

coordinate the metal ion (corresponding to *Saccharomyces cerevisiae* Asp51, His53 and His94) and the catalytic triad (*S. cerevisiae* Asp8, Lys122 and Cys167). The characteristic *cis*-peptide bond that has been identified in available nicotinamidase/pyrazynamidase structures also corresponds to conserved residues present in these species, namely Val-Ala in *Pyrococcus horikoshii, S. cerevisiae, Leishmania infantum* and *C. intestinalis* [7,38,39], Ile-Ala in *Mycobacterium tuberculosis, Acinetobacter baumanii, H. robusta* and *B. floridae* [40,41], or Val-Leu in *Streptococcus pneumoniae* [42], and are preceded by a conserved glycine that has a role in catalysis [38,40,41]. Additionally, mutations that lead to *M. tuberculosis* loss of pyrazinamidase activity have defined residues that delineate the active site scaffold [38], corresponding to *S. cerevisiae* Glu10, Asp12, Phe13, Leu20, His57, Trp91, Gly123, Tyr131, Ser132, Val162, Ala163, Tyr166 and Thr171, and most of them are conserved in all invertebrate PNC sequences as well (Figure 3B and Figure S4).

Genetic architecture conservation of *NAMPT* homologues

Given the degree of conservation of both proteins, and taking into account the divergence times of over 1000 million years between the species considered here, we have investigated the conservation at the gene structure and genome organization levels. *NAMPT* retained microsynteny in chordates, as indicated by the conserved gene order between *H. sapiens*, *M. musculus*, *D. rerio* and *B. floridae*, and also showed macrosynteny conservation in some lineages, namely between *Trichoplax adhaerens* and either *H. sapiens*, *N. vectensis* or *M. brevicollis* (Figure 4 and Figures S5-S6). For *PNC* homologues, no syntenic regions were found. Although recent studies point to a higher level of microsynteny conservation in metazoans than previously estimated [43], these evidences are challenging in some lineages due to poor genome annotation and breakdown in small scaffolds. At the level of exon-intron structure, *NAMPT* is more homogeneous than *PNC*, considering the number and size of exons, and total gene length (Figures S7-S8). The exception is observed in *N. vectensis,* where *NAMPT* has many small exons spanning 14 Kb in the genome, while *PNC* has only two exons in less than 2 Kb. Using the information on conserved motifs and gene structure, we were able to successfully identify *NAMPT* and *PNC* homologues as well as predict the corresponding gene structures in the hemichordate *S. kowaleskii*, a phylogenetic informative species (Figure S9).

Secondary structure conservation of *PNC* homologues

Nicotinamidase sequences are poorly conserved even in closely related species (Figures 2 and 3). Yet, considering some structures determined for archaea (*P. horikoshii*, PDB id: 1IM5), bacteria (*A. baumanii*, PDB id: 2WTA) and fungi (*S. cerevisiae*, PDB id: 2H0R), sharing only 30% protein identity (Figure 5A), the 3D structures are perfectly

superimposable (Figure 5B). Such structural conservation is observed across the three domains of life, as all PNC enzymes share a similar core fold (Figure S10), with a potential increase in complexity of the enzyme that is active as a monomer in *P. horikoshii* [38], dimer in *A. baumanii* [40] and heptamer in *S. cerevisiae* [39]. Thus, we have aligned PNC sequences based on secondary structure predictions and determined that invertebrate PNCs also show structural conservation (Figure 6). The regions corresponding to alpha-helices (red) and beta-sheets (yellow) are conserved at the structural level, even if the amino acids are not (Figure 6A). For instance, the alpha-helices of regions I, II and III comprise different amino acids while displaying a similar fold. To illustrate this, region II is shown in detail for *P. horikoshii*, *A. baumanii* and *S. cerevisiae* (Figure 6B).

Modeling and docking analyses of invertebrate NAMPTs and PNCs

To gain insight into the structures of invertebrate NAMPTs and PNCs, we have performed homology modeling and protein-ligand docking. To overcome limitations in the interpretation of results, we have used several templates to generate the models (Table S4). The LIGPLOT program was used to generate schematic diagrams between ligand (Nicotinamide, NCA) and receptor (NAMPT and PNC), which are shown in Figure 7. The prediction accuracy redocking test performed for the NAMPT (PDB 2E5D from *H. sapiens*) and PNC (PDB 3R2J from *L. infantum*), were in agreement with the ligand-receptor conformation in these X-ray structures. We obtained a similar active site ligand-receptor interaction for both NAMPT and PNC, which insure that the docking approach was accurate enough to be applied to the various molecular systems.

In NAMPT protein active site, all species, except *N. vectensis,* maintained most of the ligand-receptor interactions when compared with the structure of human NAMPT (Figure 7A). The homologous NAMPT of *B. floridae* has a hydrogen bond network that stabilizes the active site with two H-bonds between the side-chain of Arg-293 and the oxygen atom of the ligand. A similar bonding network can be observed in the human protein (PDB 2E5D) where Asp-219 binds to the nitrogen atom of the substrate (NCA). Hydrophobic interactions are similar when compared with the human active site. In *C. teleta*, H-bond interactions between Arg-300 and NCA oxygen moiety and between Asp-209 and Asp-16 to both NCA nitrogens preserve the NCA conformation in the active site. Two hydrophobic interactions in *C. teleta* (Tyr-18 and Phe-183) with ligand atoms are not seen. In *N. vectensis* no H-bond interaction is present, but the most important hydrophobic interactions, Phe-182(B), Arg-298(B) and Tyr-17(A), are preserved. The H-bond interaction network of *S. purpuratus* shows that Asp-210(B) H-bond is maintained. Two other H-bonds, Tyr-19(A) and Glu-235(B), and hydrophobic interactions of the NCA ligand to Phe-184 (B) and Ala-233 (B) are also present. Globally, the NAMPT binding modes obtained by docking for the species analysed shared

the critical hydrophobic and hydrogen bonding interactions and, if not (e.g. *N. vectensis*), the conformational status of NCA was maintained.

Next we also analysed the conformational changes of PNC active and catalytic sites (flexible residues) in the four species (Figure 7B). In the *B. floridae* PNC, Phe-22, Trp-110, Val-182 and Cys-183 hydrophobic interactions contribute to the binding status of NCA. The three hydrogen bonding interactions (His-113 to NCA oxygen atom, Asp-62 to NCA nitrogen atom and His-113 to Asp-62) sustain the conformational position of the ligand. The Zn^{2+} keeps the strong binding to the ligand that was also present in *L. infantum* PNC (PDB 3R2J). In *C. teleta* the Trp-110 (hydrophobic interaction) and Tyr-147 (H-bond interaction) are the residues from the active site that play an important role in the ligand-receptor interaction. As in *B. floridae*, the His-113 has a hydrogen bond connection with NCA. Two other H-bond interactions not present in *L. infantum* PNC (Ser-70 and Lys-108) appear to be important to ligand binding. The interaction between Zn^{2+} and ligand is maintained. Although no significant changes in ligand conformation were observed, hydrogen bonds in *N. vectensis* were not predicted. When compared to 3R2J, hydrophobic interactions Cys-21, Trp-99, Ala-169 and Cys-173 are kept for the active site residues. Hydrophobic interactions for the catalytic residue Cys-173 are present, as well as a newly arisen Phe-11 interaction with the ligand. In *S. purpuratus*, hydrophobic contacts between Tyr-106, Trp-143, Ala-175 and the ligand are retained. Catalytic site residues Asp-17 and Cys-179 also bind to the ligand through an H-bond and a hydrophobic interaction. Two unique hydrogen bonds (Asp-57 and Leu-174) and hydrophobic contacts (His-109 and Phe-22) arise in the ligand-protein interaction. It can also be noticed that a conserved histidine (His-113 in *B. floridae* and *C. teleta*, and His-109 in *S. purpuratus*) maintains the interaction with the ligand.

Discussion

Nicotinamide phosphoribosyltransferases (NAMPTs) and nicotinamidases (PNCs) are the main NAD salvage enzymes and, until recently, were thought to occur in distinct lineages. Our data show that several Metazoa species have predicted homologues of both enzymes and that both genes are simultaneously expressed in *B. floridae, S. purpuratus*, *C. teleta* and *N. vectensis*. The distribution of NAMPT and PNC homologues points to the presence of both genes in early eukaryote evolution with selective gene loss and retention in different animal lineages. Interestingly, loss of either one of the genes was predominantly found in fast evolving lineages, namely *D. melanogaster*, *C. elegans* and *C. intestinalis*, while slow-evolving species such as *B. floridae* retained both [44]. This is also reflected in genome architecture, with conserved *NAMPT* microsynteny in vertebrates and *B. floridae*.

We also highlight different conservation patterns in NAMPT and PNC homologues, at the protein amino acid sequence and at the 3D structural level. NAMPT sequences are highly conserved, as evidenced by small evolutionary divergences between species and long stretches of identical amino acids surrounding important catalytic and structural positions. As dimerization is required for NAMPT activity [25], in addition to active site residues that interact with the substrates and reaction products, amino acids that constitute the dimer interface are also conserved. For PNCs the sequence identity is lower, yet, critical amino acids are conserved and the overall fold is maintained in all the three domains of life. These are unifying features of nicotinamidases, even though there is a diversity of catalytic mechanisms described, with some exceptions concerning metal binding and metal ion coordination [7,20,21,29,41,42].

Homology modeling and protein-ligand docking indicates that active site residues and interactions of invertebrate NAMPTs with the substrate, nicotinamide, are similar to what is described for vertebrate NAMPTs [24,25,35-37]. In invertebrate PNCs, most interactions are maintained while additional hydrogen bonds and hydrophobic contacts were found. These new interactions might derive from complementary amino acid changes as a result of epistatic interactions between residues [21,45], which is consistent with a structural conservation of PNCs.

Our analyses validate invertebrate NAMPTs and PNCs, suggesting that both the two-step and the four-step NAD salvage pathways are functional in these organisms. These findings imply that either these enzymes are not redundant, or that specific metabolic requirements call for increased NAD production in some species that only the presence of both enzymes would fulfil.

Materials and Methods

Sequence analysis

The human *NAMPT* and the yeast *PNC1* amino acid sequences were used as queries in BLAST searches [46], from National Center of Biotechnology Information, NCBI (http://www.ncbi.nlm.nih.gov/sites/genome) and Joint Genome Institute, JGI (http://genome.jgi-psf.org/) sequenced genomes. In organisms with multiple hits, the reciprocal best hit was selected for further analysis. All sequences retrieved in this process and further analysed are listed in Table S1. Estimates of evolutionary divergence between sequences were conducted in MEGA5 [47] and calculated as the number of amino acid substitutions per site. Analyses were conducted using the Poisson correction model [48] and involved 13 amino acid sequence homologues for each protein. Positions containing gaps

and missing data were eliminated, resulting in a total of 436 (NAMPT) and 167 (PNC) positions in the final dataset. Alignments were visualized in Geneious [49] v5.5.6 to generate logos. Structural alignments of PNC homologues were performed in Ali2D (http://toolkit.tuebingen.mpg.de/ali2d). Divergence times between species were estimated using Time Tree (http://www.timetree.org/) [50]. MATLAB version R2010b was used to generate 3D graphics (the input data is shown as Table S3) and calculate Kendall rank correlation coefficients. Correlations were measured against a reference function consisting of a monotonic increasing function of protein distances against evolutionary divergence (the hypothesis). Synteny was determined using the CHSminer software (http://www.biosino.org/papers/CHSMiner/) [51] and the JGI genome portal (http://genome.jgi-psf.org/). *Saccoglossus kowaleskii* BLAST searches were also performed as described above (http://blast.hgsc.bcm.tmc.edu/blast.hgsc?organism=20), the corresponding genome contigs (115790 and 40985) were retrieved and the NAMPT and PNC protein sequences were manually predicted, based on the conserved motifs identified. Exon predictions were then performed in Genescan (http://genes.mit.edu/GENSCAN.html).

Molecular homology modeling

Prime [52] was used to search homologous proteins in NCBI PDB database (http://www.rcsb.org/pdb/home/home.do) for PNC and NAMPT. PDB templates (Table S4) were selected considering lowest e-values (<1x10-6), and structures without many missing residues (gaps<20%). PNC and NAMPT sequences for the species *B. floridae*, *C. teleta*, *S. purpuratus* and *N. vectensis* were used to generate the alignments with homologue proteins. For secondary structure prediction the third-party program SSpro [53,54] was used and then all the templates re-aligned to the query sequence. The resulting alignment was used to build the protein models. The LigPrep [55] interactive optimizer (protein preparation wizard) with neutral pH was used to optimize the protein model. Finally hydrogens were added, bond order was assigned and selenomethionines were converted to methionines for the generated models.

Molecular docking simulations

The 3D-structures of ligands were obtained from the PDB structures. The protein-ligand complexes were prepared with AutoDockTools [56,57]: hydrogen atoms were added for each protein and Kollman united atom charges assigned. Hydrogens were also added to the ligand (NCA) and charges were calculated by the Gasteiger-Marsili method. The rotatable bonds in the ligands were assigned with AutoDockTools. The Zn atom of PNC was assigned a charge of +2. AutoDock4.2 [56] was used to perform protein-ligand docking calculations. To insure the accuracy of our methodological approach we first have done

redocking of the two most recently available X-ray structures (NAMPT PDB 2E5D and PNC PDB 3R2J) and then applied it to the various predicted protein-ligand systems. Various grid sizes were tested using as structural criteria the similarity between our docked results and the X-ray structure of *H. sapiens* NAMPT (2E5D) and *L. infantum* PNC (3R2J). We have selected a cubic grid box of 30×25×40 Å for NAMPT and 35x35x40 Å for PNC, centered on the C2-C5 ligand atoms distance mean with a grid spacing of 0.375 Å as shown in Table S4.

We considered the binding pockets described in the literature [7,36] (also shown in Table S5) to perform the flexible protein-ligand docking. The corresponding residues in the homology alignment are described in Table S5. We performed the docking simulations using 100 independent Lamarckian genetic algorithm (LGA) runs, with the population size set to 200, the number of energy evaluations set to 10 000 000 and the maximum number of generations set to 27 000. All other parameters were used as default [58,59]. The results were analysed clustering together the conformations within a RMSD of 2 Å. The cluster with lower energy and with a conformation similar to the X-ray structure of NAMPT (PDB id: 2E5D) and PNC (PDB id: 3R2J) was selected for each species.

H-bonds and hydrophobic interactions for ligand-receptor molecules

Interactions between the ligand (NCA) and receptors (NAMPT and PNC) were calculated using LIGPLOT [60]. The hydrogen bonds were calculated using geometrical criteria [61] of protein-ligand complex (The used criteria is: H–A distance <2.7 Å, D–A distance <3.3 Å, D–H–A angle >90°, D–A–AA angle >90° and H–A–AA angle >90°, where A is the hydrogen acceptor, D is the hydrogen donor, AA is the atom attached to the hydrogen acceptor, and H an atom of hydrogen). LIGPLOT also calculates non-covalent bond interactions (hydrophobic interactions) by applying a simple cut-off of 3.9 Å. LIGPLOT diagrams were generated for each species. PyMOL [62] was used to generate the 3D images.

Expression analysis

B. floridae (whole organism), *C. teleta* (whole organism), *S. purpuratus* (gonad) and *N. vectensis* (whole organism) samples were obtained from Ocean Genome Legacy (OGL Accession ID numbers S13045, S13061, S13034 and S13115, respectively) [63]. RNA was extracted with the Illustra TriplePrep kit (GE Healthcare) and genomic DNA was removed from RNA preparations with an additional DNase treatment using DNase I, RNase-free (Fermentas, Thermo Fisher Scientific Inc.), according to the manufacturer's procedure. Complementary DNA (cDNA) was synthesized from 1µg of total RNA using the RETROscrip® First Strand Synthesis Kit (Ambion) with oligo-dT primers according to the manufacturer's instructions. Reverse-transcription PCR reactions were prepared using

HotStarTaq® Master Mix Kit (Qiagen) with 2µl of the synthesized cDNA in a 10µl final volume. Q solution was included in the reaction (10%) in NAMPT amplification in *B. floridae* and *S. purpuratus*. PNC and NAMPT were amplified with species-specific primers described in Table S6, with a final concentration of 0.2 µM. Thermocycling conditions were as follows: initial denaturation at 95°C for 15min, 40 cycles at 95ºC for 30sec, variable annealing temperatures ranging from 52ºC to 62ºC (Table S6) for 1min30sec, and 72°C for 1min, and a final extension step of 10min at 72°C. All amplification products were visualized on 1.5% agarose gels and were confirmed by sequencing. For that, PCR products were purified with ExoSAP-IT (USB Corporation) by incubation at 37°C for 15 min, followed by enzyme inactivation for 15 min at 85°C. The resulting purified fragments were sequenced using an ABI Big Dye Terminator Cycle Sequencing Ready Reaction kit v 3.1 (Applied Biosystems) and analysed in an ABI PRISM 3130xl (Applied Biosystems).

Expressed Sequence Tag (EST) information was retrieved from available databases for *B. floridae* [64]*, S. purpuratus* [65] and *N. vectensis* [66] and is detailed in Table S2.

Acknowledgments

We are thankful to Fátima Gil from Aquário Vasco da Gama (Lisboa, Portugal), Ana Lopes Ribeiro from INEB (Porto, Portugal) and Timery S. DeBoer from Ocean Genome Legacy (Ipswich MA, USA) for providing samples.

Figures

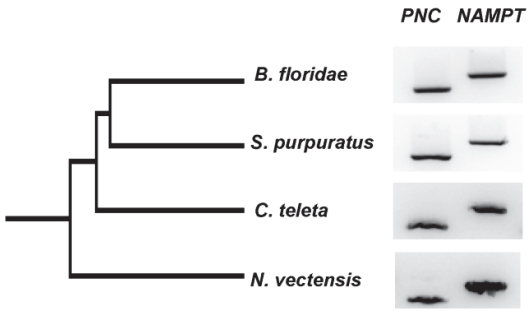


Figure 1: NAMPT and PNC homologues are co-expressed in invertebrates. RT-PCR analysis shows that NAMPT and PNC are simultaneously expressed in *Branchiostoma floridae*, *Strongylocentrotus purpuratus*, *Capitella teleta* and *Nematostella vectensis*.

	Hs	Mm	Dr	Bf	Ci	Sp	Ct	Hr	Lg	Dm	Ce	Nv	Ta	Sc	Mb
H. sapiens		-	-	-	-	-	-	-	-	-	-	-	-	-	-
M. musculus	0,035		-	-	-	-	-	-	-	-	-	-	-	-	-
D. rerio	0,119	0,114		-	-	-	-	-	-	-	-	-	-	-	-
B. floridae	0,483	0,498	0,494		0,618	0,607	0,513	0,855	0,944	0,664	0,841	0,533	-	0,898	0,749
C. intestinalis	-	-	-	-		0,687	0,749	0,898	0,991	0,749	0,898	0,629	-	0,869	0,749
S. purpuratus	0,509	0,509	0,505	0,341	-		0,736	0,787	0,898	0,787	0,944	0,736	-	1,007	0,724
C. teleta	0,532	0,525	0,513	0,374	-	0,367		0,913	0,944	0,585	0,913	0,675	-	1,058	0,787
H. robusta	0,589	0,58	0,597	0,436	-	0,494	0,443		0,841	1,007	1,007	0,828	-	1,04	0,814
L. gigantea	0,564	0,556	0,552	0,371	-	0,394	0,335	0,454		1,075	1,007	0,855	-	1,04	0,991
D. melanogaster	-	-	-	-	-	-	-	-	-		0,991	0,675	-	1,007	0,652
C. elegans	-	-	-	-	-	-	-	-	-	-		0,801	-	1,075	0,959
N. vectensis	0,505	0,502	0,49	0,325	-	0,397	0,312	0,502	0,364	-	-		-	0,869	0,749
T. adhaerens	0,548	0,56	0,544	0,377	-	0,443	0,367	0,505	0,415	-	-	0,377		-	-
S. cerevisiae	-	-	-	-	-	-	-	-	-	-	-	-	-		1,093
M. brevicollis	0,576	0,576	0,556	0,367	-	0,45	0,415	0,505	0,461	-	-	0,408	0,454	-	

Figure 2: Evolutionary divergence between NAMPT and PNC homologues. The estimates of evolutionary divergence were calculated as amino acid substitutions per site between NAMPT (green) and PNC (orange) sequences for several species. Hs, *Homo sapiens*; Mm, *Mus musculus*; Dr, *Danio rerio*; Bf, *Branchiostoma floridae*; Ci, *Ciona intestinalis*; Sp, *Strongylocentrotus purpuratus*; Ct, *Capitella teleta*; Hr, *Helobdella robusta*; Lg, *Lottia gigantea*; Dm, *Drosophila melanogaster*; Ce, *Caenorhabditis elegans*; Nv, *Nematostella vectensis*; Ta, *Trichoplax adhaerens*; Sc, *Saccharomyces cerevisiae*; Mb, *Monosiga brevicollis*.

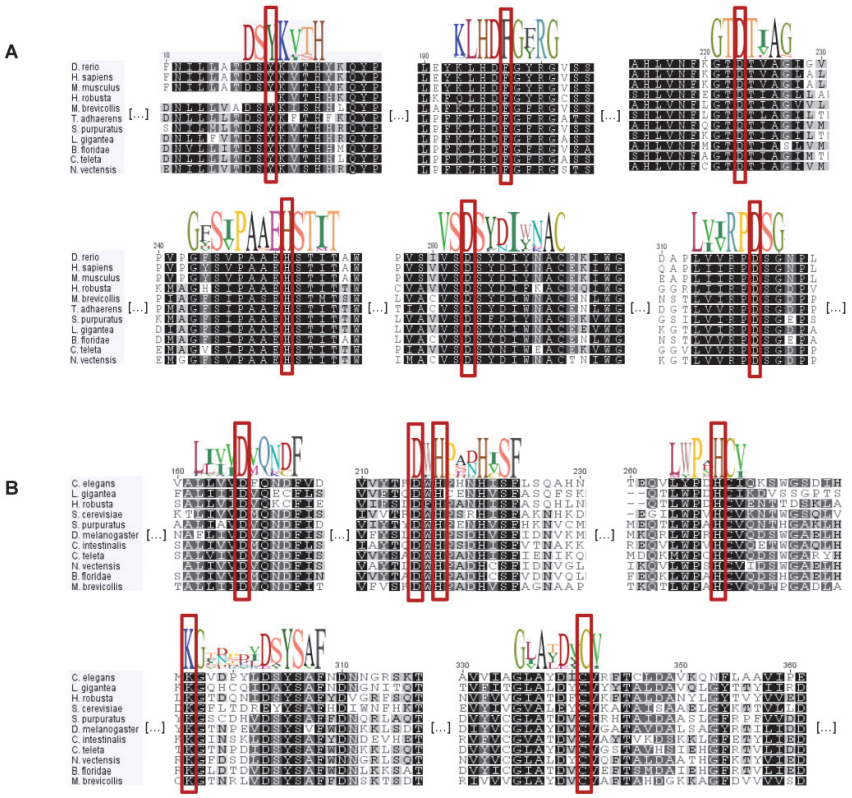


Figure 3: Amino acid motifs found in NAMPT and PNC homologues. The conserved amino acid motifs surrounding the active site residues (boxed) are shown as logos and displayed above the aligned sequences. NAMPT conservation is highlighted by the large blocks of identical amino acids that are found in the species analysed **(A)**. In PNC homologues, although the overall amino acid identity is low, the presence of conserved motifs is still detected throughout the species analysed that range wide evolutionary distances **(B)**.

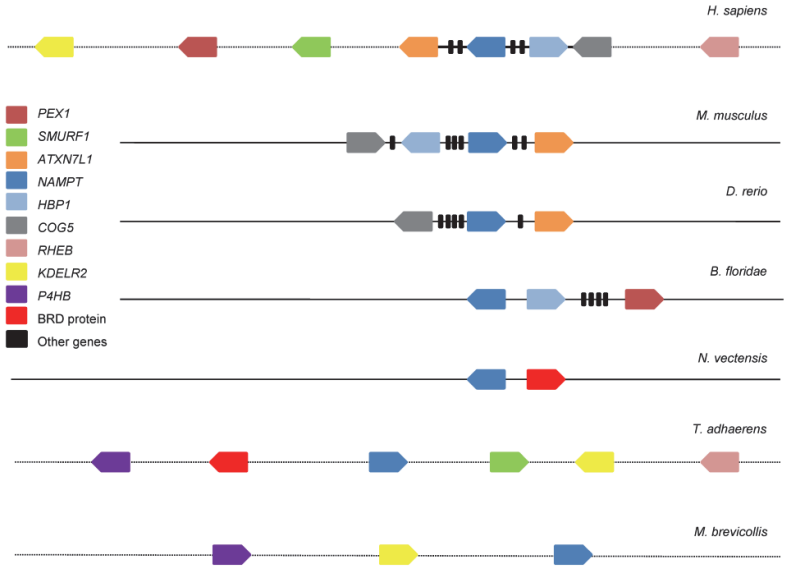


Figure 4: Syntenic organization of NAMPT homologues. Gene order and organization are represented for several lineages, and show conservation of microsynteny in chordates. *H. sapiens* chromosome 7, *M. musculus* chromosome 12, *D. rerio* chromosome 4, *B. floridae* scaffold 633, *N. vectensis* scaffold 360, *T. adhaerens* scaffold 2 and *M. brevicollis* scaffold 7 are displayed and dots indicate intervals containing multiple genes (>4).

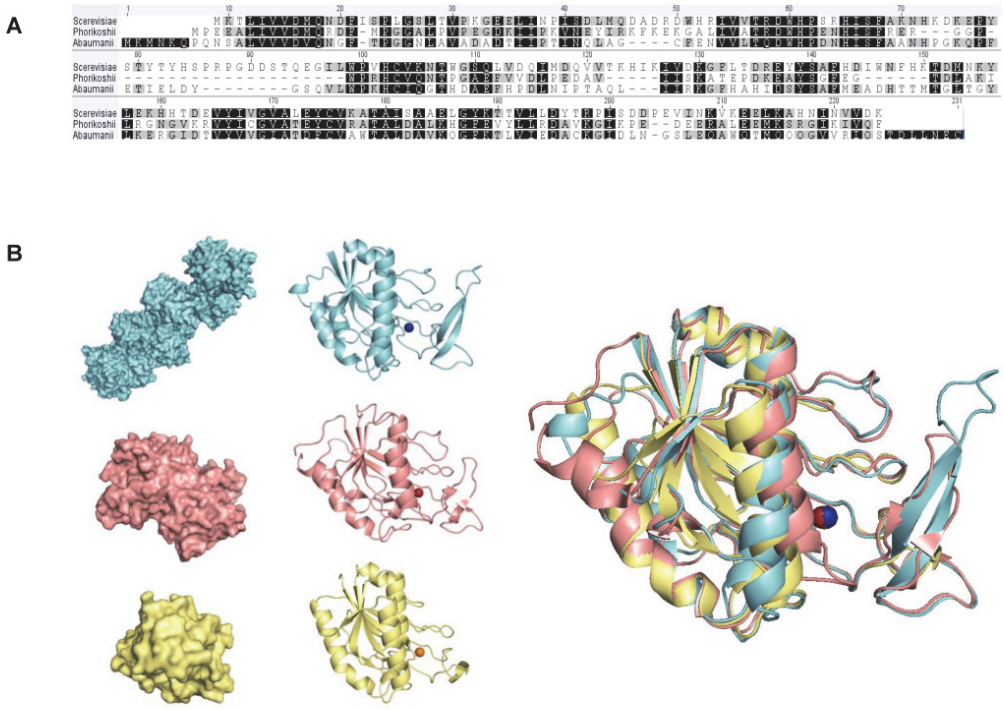


Figure 5: Structural conservation between PNC homologues. Alignment of sequences **(A)** and structures **(B)** of PNC homologues from *P.horikoshii* (yellow), *A.baumanii* (pink) and *S.cerevisiae* (blue). Although there is an increasing structural complexity from a monomer in Archaea, a dimer in Bacteria and an heptamer in Fungi and the amino acid identity of the sequences is around 30%, the 3D structural subunits of PNC homologues are superimposable.

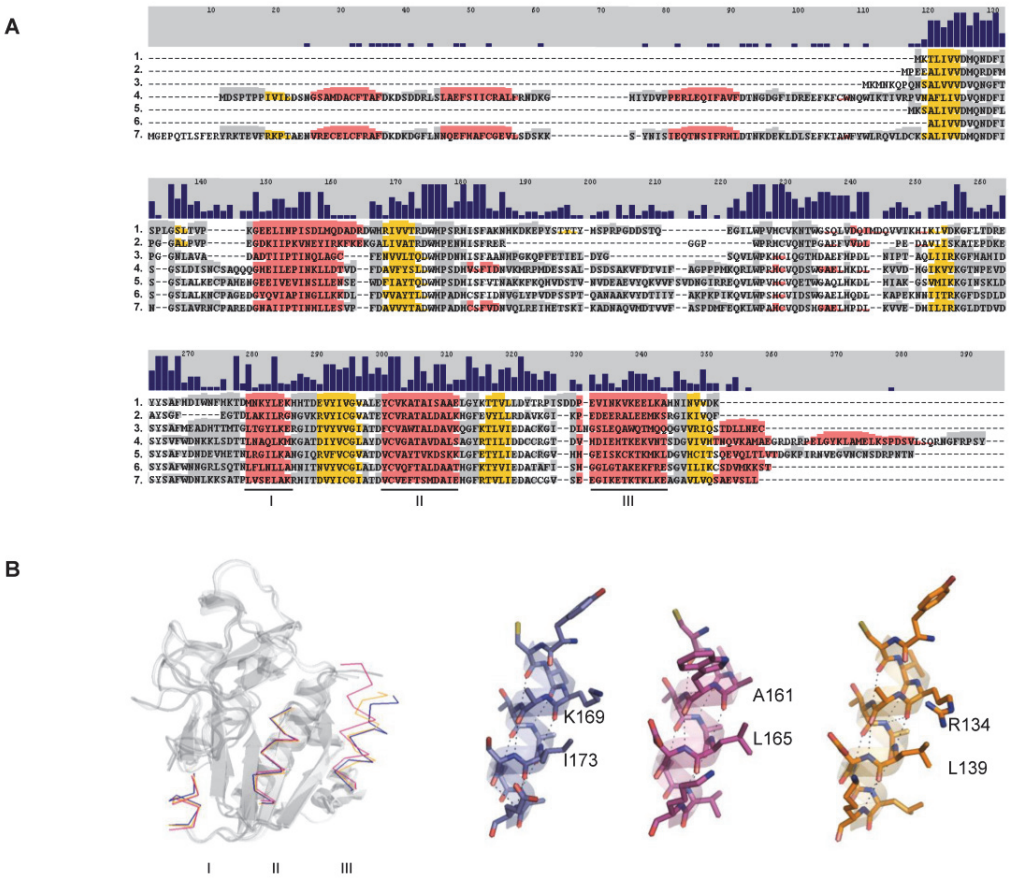


Figure 6: Predicted secondary structure of PNC homologues. (A) Aligned amino acid sequences of representative PNC homologues are displayed in function of the secondary structure. Alpha-helices are shown in red, beta-sheets are in yellow and grey represents coiled regions. Regions of structural conservation are highlighted in colour even when the primary sequences are not conserved as demonstrated by the graphic bars above the sequences. 1, *Saccharomyces cerevisiae*; 2, *Pyrococcus horikoshii*; 3, *Acinetobacter baumanii*; 4, *Drosophila melanogaster*; 5, *Ciona intestinalis*; 6, *Nematostella vectensis*; 7, *Branchiostoma floridae*. **(B)** Alpha-helices I, II and III formed by groups of unrelated amino acids are structurally equivalent as shown by the 3D superimposition. In blue, *S. cerevisiae*; in pink, *A.baumanii*; and in yellow, *P. horikoshii*.

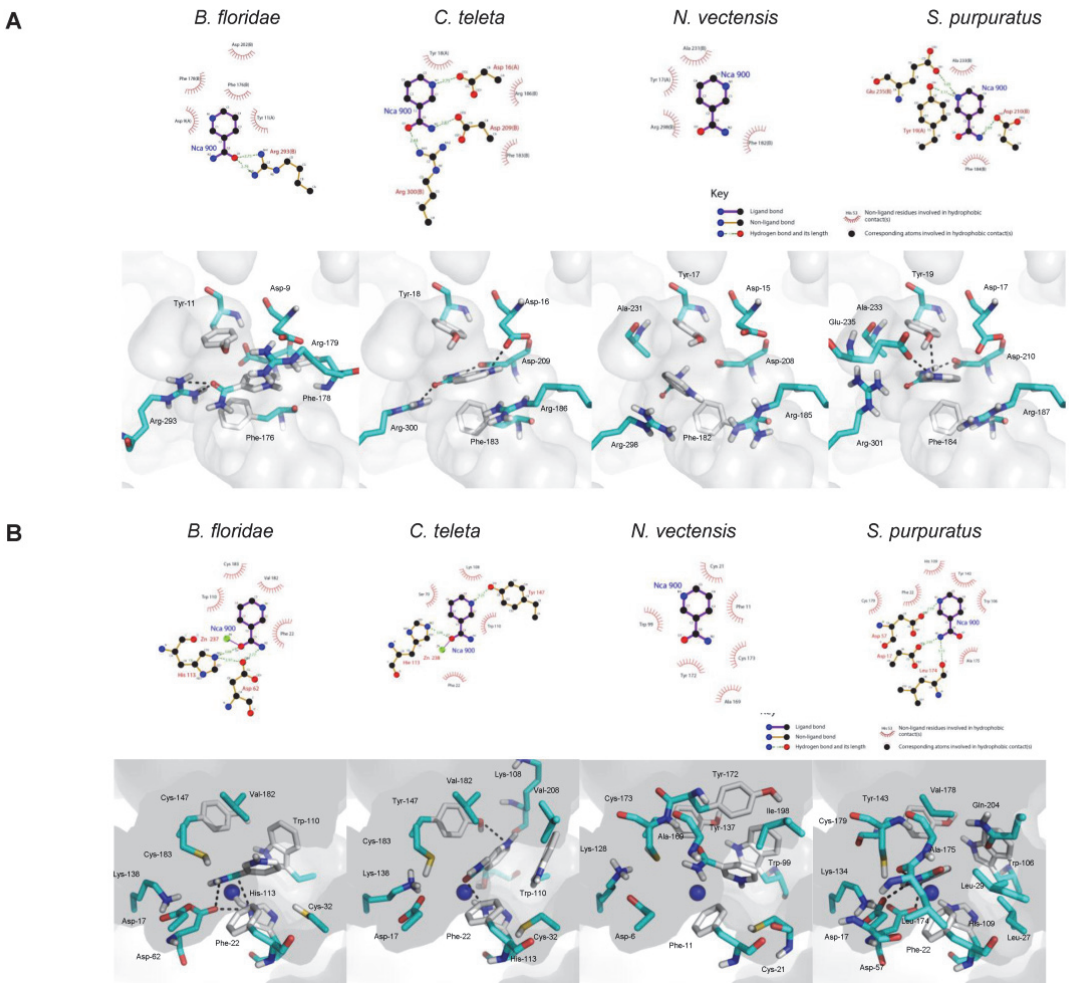


Figure 7: Hydrophobic interactions and hydrogen bonding network between the ligand (NCA) and the various receptors. NAMPT (A) and PNC (B) representations in LIGPLOT (upper panels) and PyMOL (lower panels) representations are shown for *Branchiostoma floridae*, *Capitella teleta*, *Nematostella vectensis*, and *Strongylocentrotus purpuratus*. The major binding determinants are represented in cyan stick. The Zn^{2+} atom is in blue vdW representation.

Supporting Information

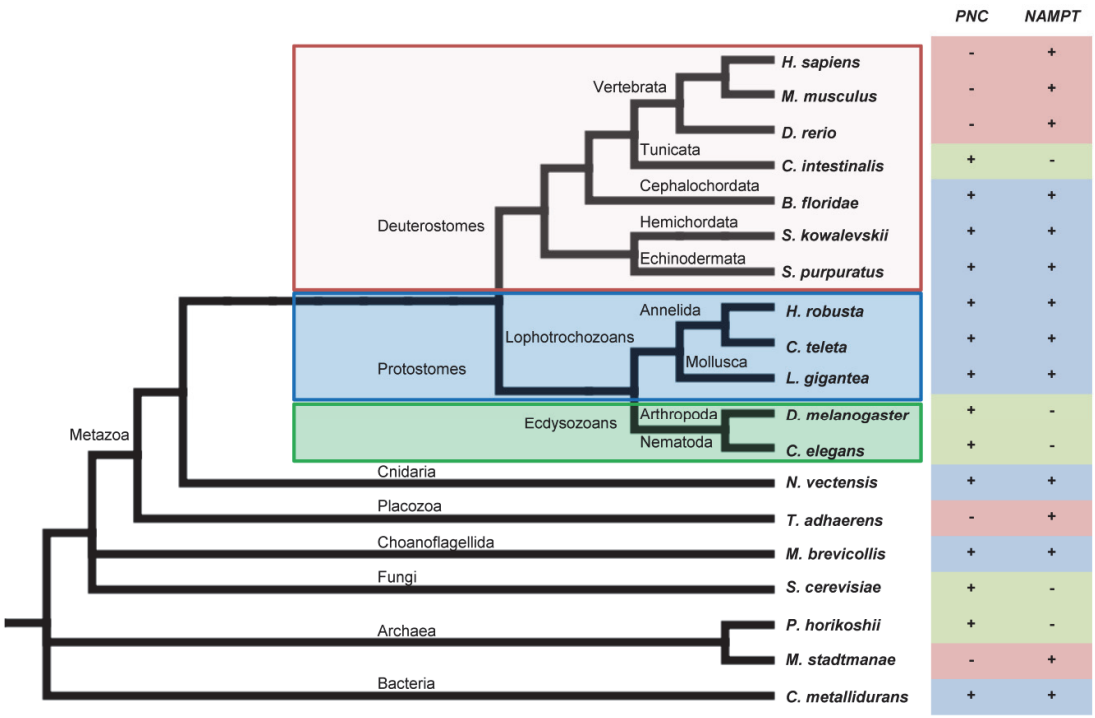


Figure S1. Distribution of *NAMPT* and *PNC* homologues across the tree of life. The presence or absence of *NAMPT* and *PNC* sequences is indicated on the columns by a plus or a minus. Protostomes are divided in ecdysozoans and lophotrochozoans (green and blue boxes of the tree, respectively), while Deuterostomes are represented in the red box.

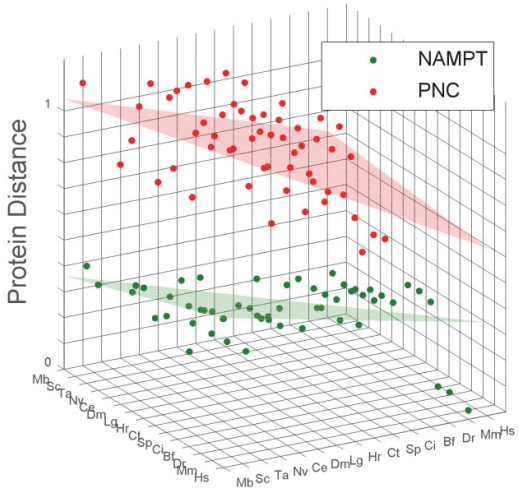


Figure S2. Evolutionary divergence between NAMPT and PNC homologues. Protein distances were plotted for each pair of species arranged accordingly to their respective divergence time. This plot shows that NAMPT is highly conserved across large evolutionary distances, while PNC is less conserved even in closely related species. Notice that in addition to being highly conserved, protein distances and evolutionary distances are correlated in NAMPT (quantified by the Kendall coefficient of 0.413), as opposed to PNC (where the Kendall coefficient was -0.052).

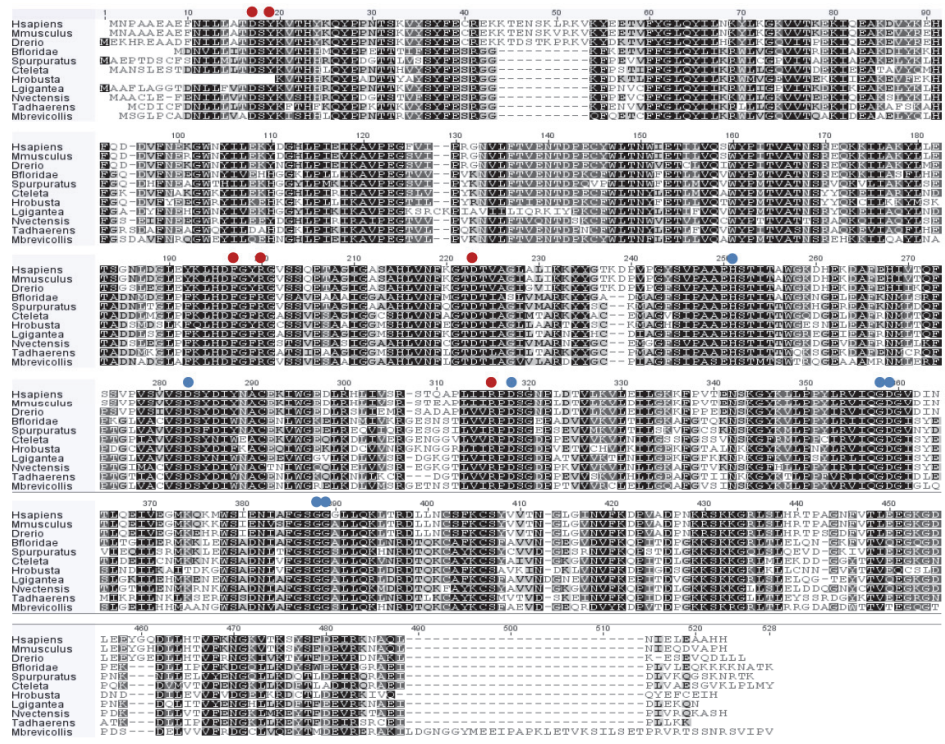


Figure S3. Alignment of the amino acid sequences from *NAMPT* homologues. Catalytic residues are marked with red dots and residues that bind nicotinamide, ribose, phosphate or NMN are highlighted in blue.

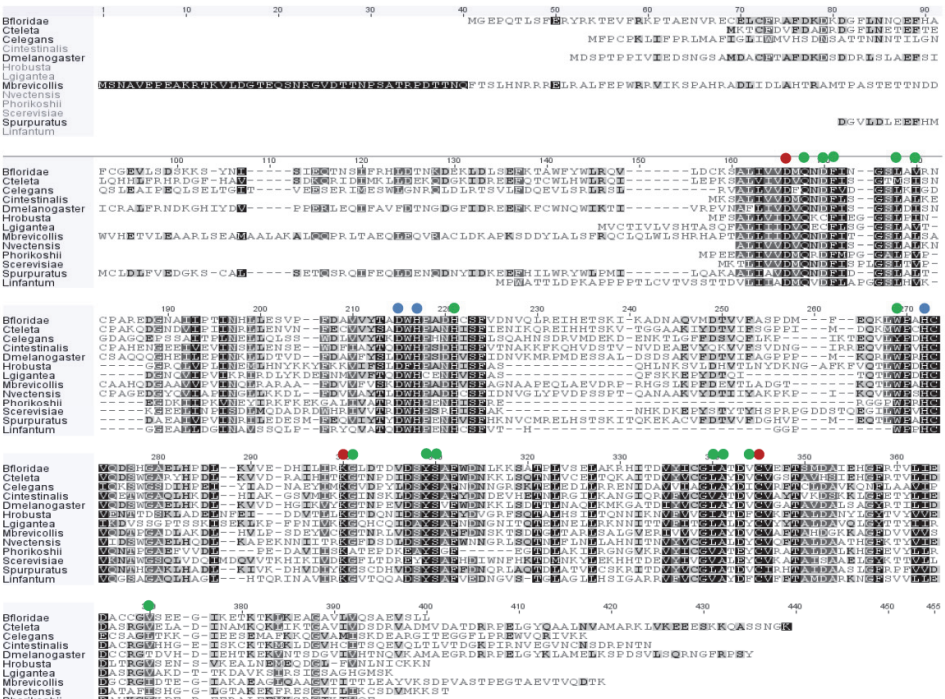


Figure S4. Alignment of the amino acid sequences from *PNC* homologues. Catalytic residues are marked with red dots and residues that bind zinc are highlighted in blue. Additional residues of the active site are shown in green.

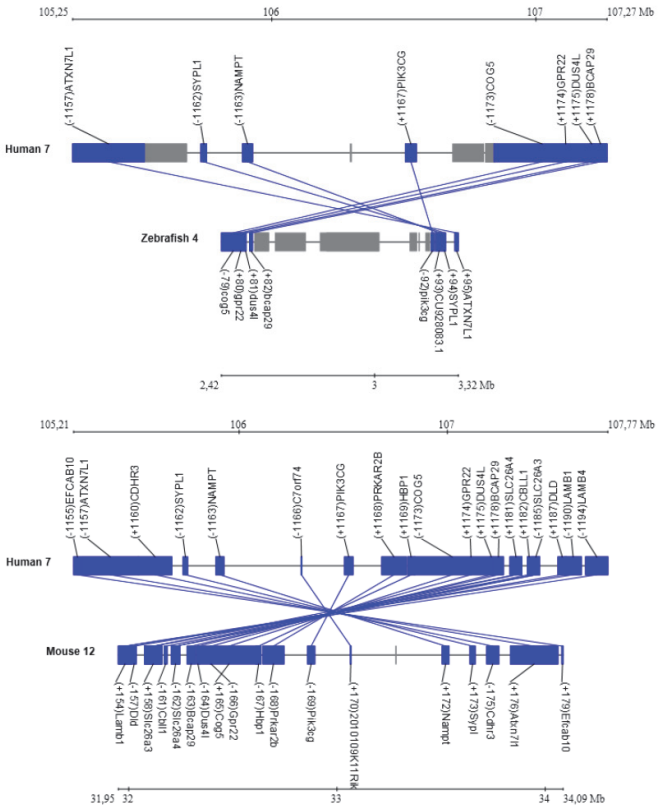


Figure S5. Vertebrate *NAMPT* synteny. Conserved synteny blocks detected between the Human, Mouse and Zebrafish genomes. Input data was automatically retrieved from Ensembl release 64 using CHSminer. Corresponding chromosomes are indicated.

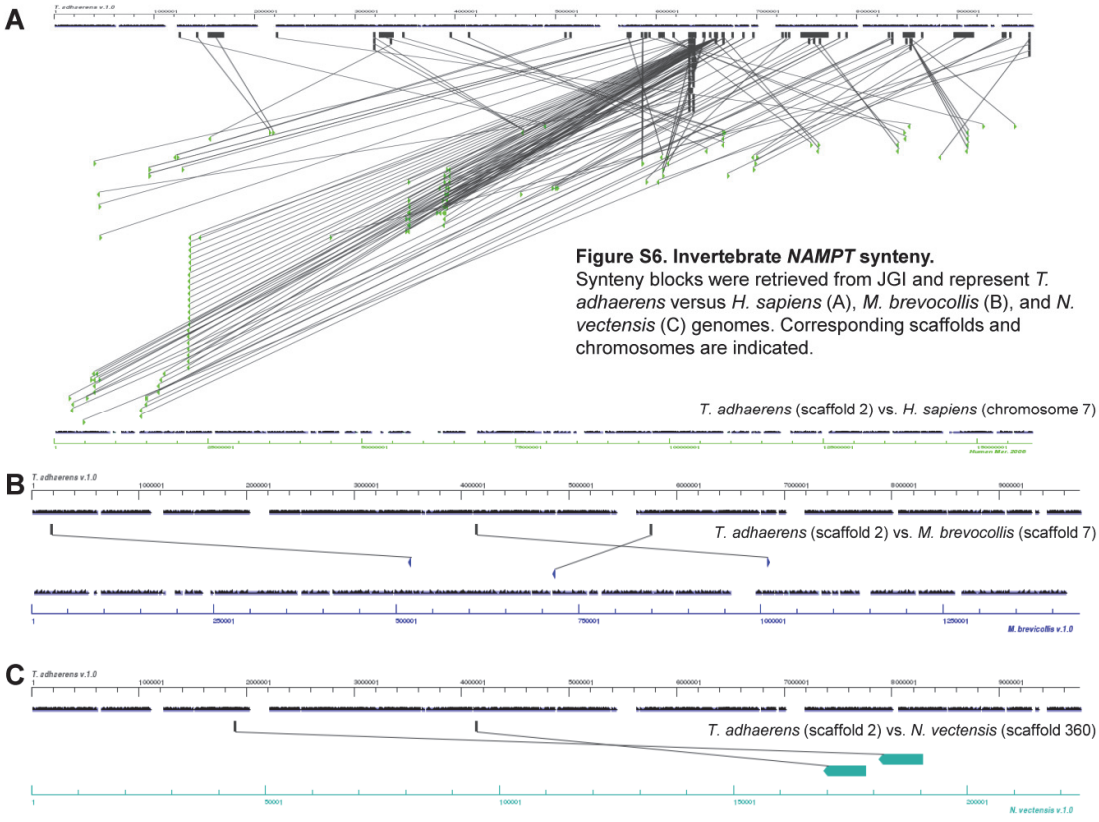


Figure S6. Invertebrate *NAMPT* synteny. Synteny blocks were retrieved from JGI and represent *T. adhaerens* versus *H. sapiens* (A), *M. brevocollis* (B), and *N. vectensis* (C) genomes. Corresponding scaffolds and chromosomes are indicated.

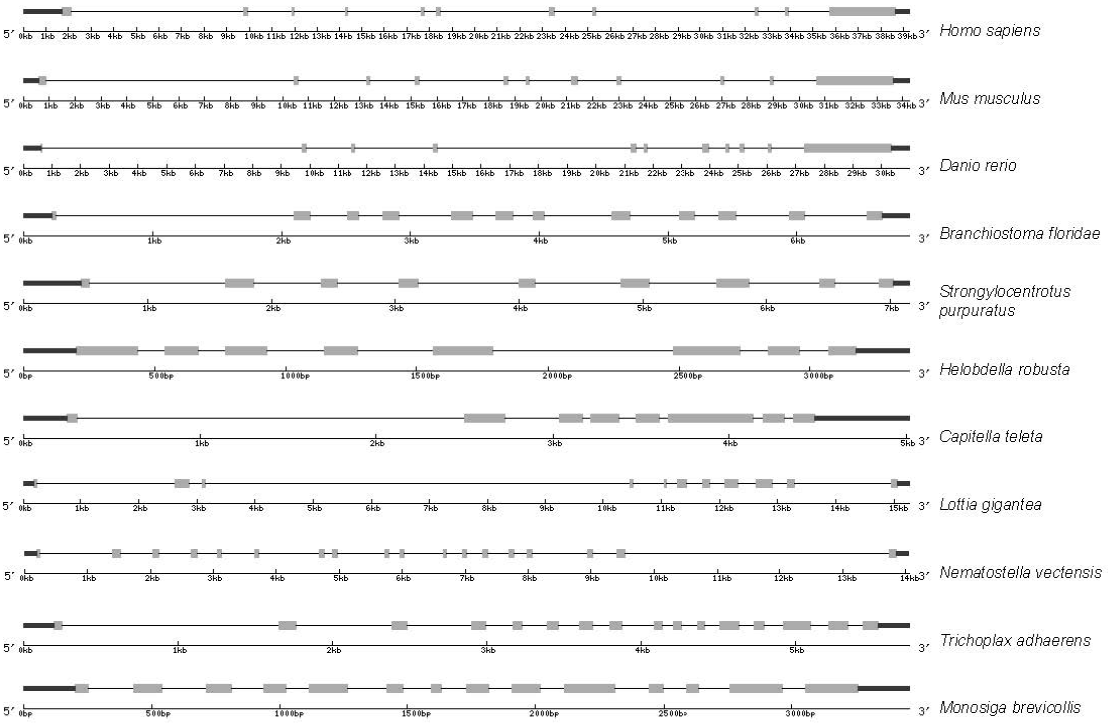
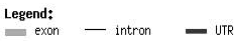


Figure S7. *NAMPT* gene structure.
Exon-intron predictions were performed in Gene Structure Display Server (http://gsds.cbi.pku.edu.cn/).



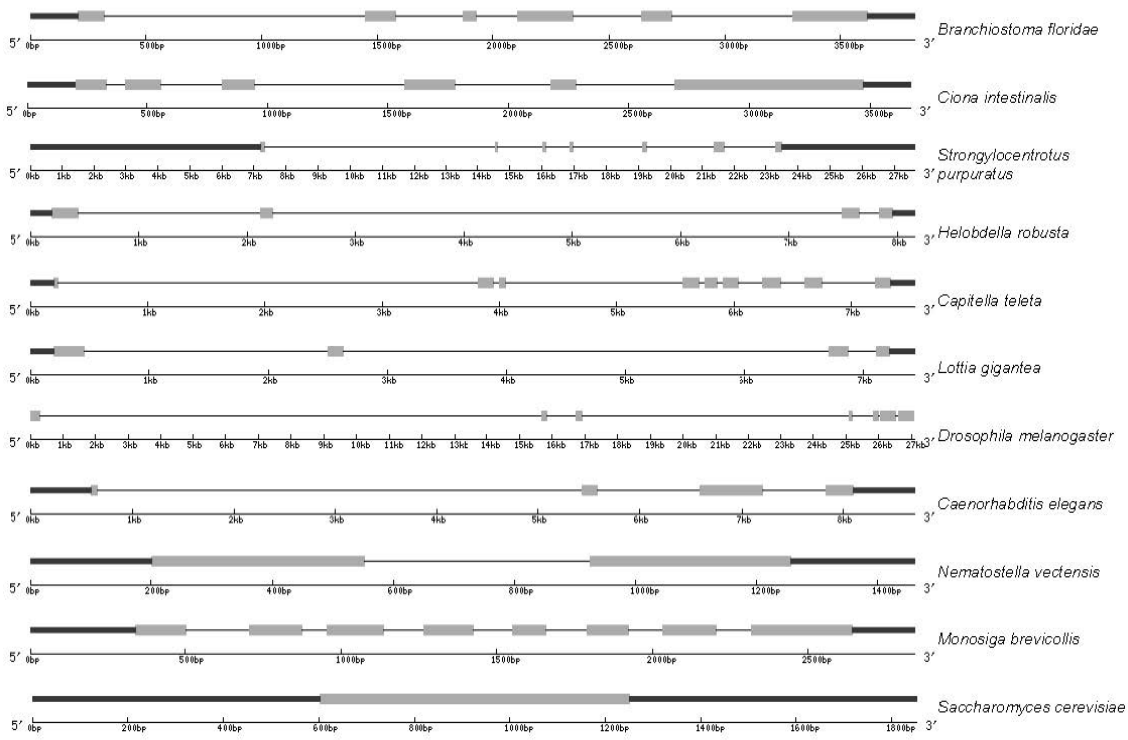
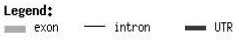


Figure S8. *PNC* gene structure.
Exon-intron predictions were performed in Gene Structure Display Server (http://gsds.cbi.pku.edu.cn/).



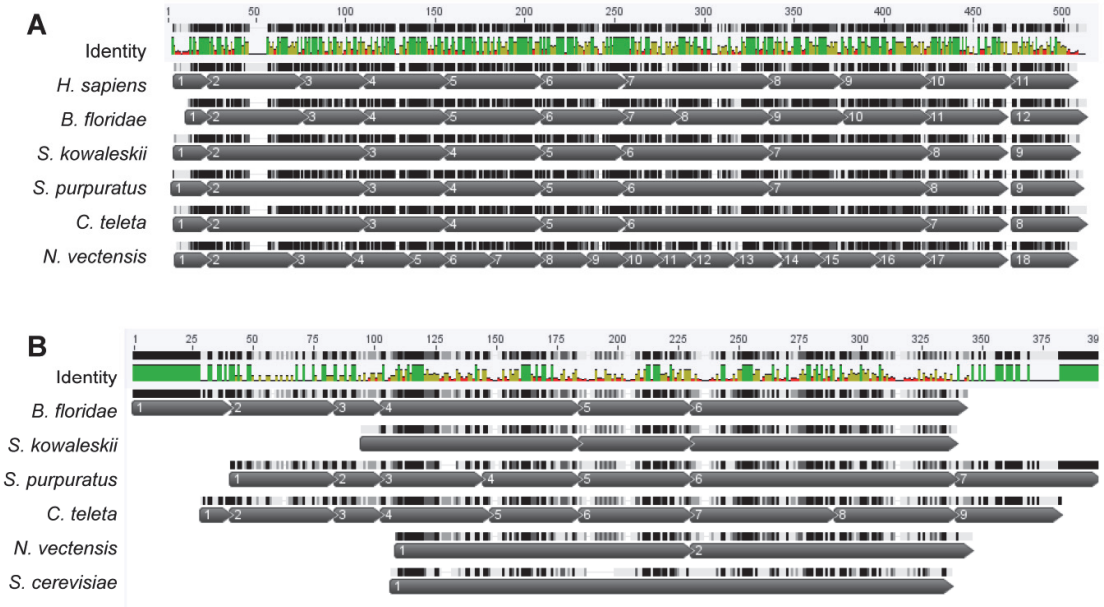
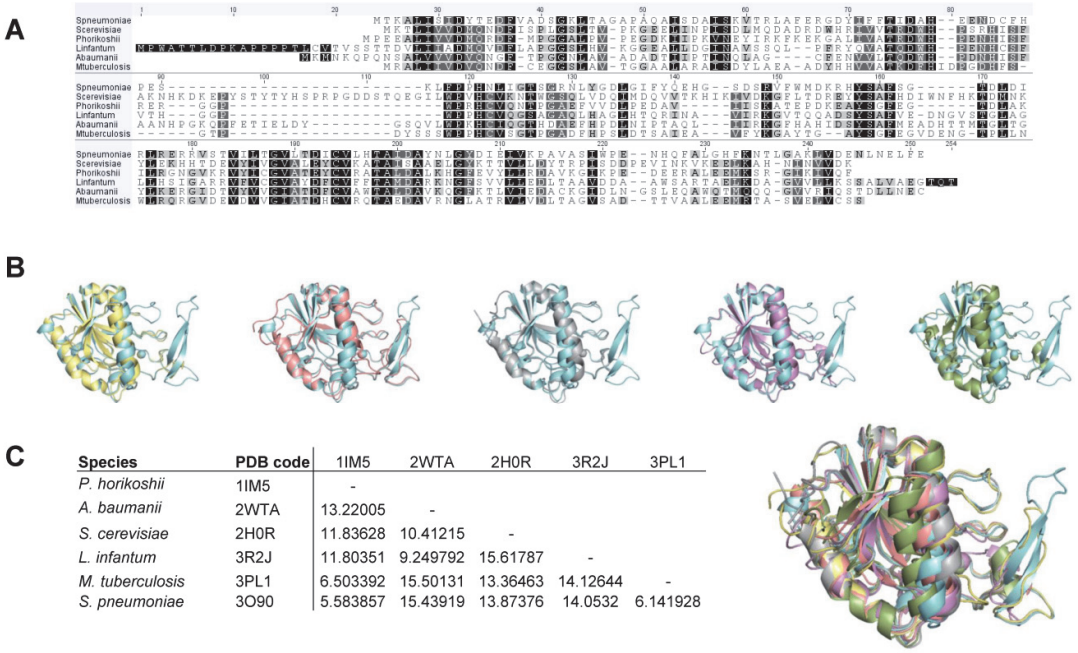


Figure S9. *NAMPT* (A) and *PNC* (B) amino acid alignments indicating exon size and number. *Saccoglossus kowaleskii* (Sk) protein sequences were manually predicted, based on the conserved motifs identified in this paper. Exon predictions were performed in Genescan (http://genes.mit.edu/GENSCAN.html).



Species	PDB code	1IM5	2WTA	2H0R	3R2J	3PL1
P. horikoshii	1IM5	-				
A. baumanii	2WTA	13.22005	-			
S. cerevisiae	2H0R	11.83628	10.41215	-		
L. infantum	3R2J	11.80351	9.249792	15.61787	-	
M. tuberculosis	3PL1	6.503392	15.50131	13.36463	14.12644	-
S. pneumoniae	3O90	5.583857	15.43919	13.87376	14.0532	6.141928

Figure S10. *PNC* alignments of all the templates available. (A) Alignments of the amino acid sequences. (B) Structural alignments of 2H0R (blue), 1IM5 (yellow), 2WTA (pink), 3R2J (grey), 3PL1 (purple) and 3O90 (green). (C) RMSD scores of the structural alignments. All structures are superimposed on the right.

Supplementary Table S1: Sequences of NAMPT and PNC orthologues.

Species	Genome source	PNC	NAMPT
Cupriavidus metallidurans	NCBI	YP_584002.1	YP_587233.1
Pyrococcus horikoshii	NCBI	NP_142913.1	-
Metanosphaera stadtmanae	NCBI	-	ABC57157.1
Saccharomyces cerevisiae	EnsEMBL	PNC1 (YGL037C)	-
Monosiga brevicollis	JGI	33390	24677
Trichoplax adhaerens	JGI	-	20412
Nematostella vectensis	JGI	94959	135670
Caenorhabditis elegans	EnsEMBL	Y38C1AA.3a	-
Drosophila melanogaster	NCBI	NP_732446.1	-
Lottia gigantea	JGI	126851	218342
Capitella teleta	JGI	144642	162451
Helobdella robusta	JGI	71646	67798
Strongylocentrotus purpuratus	NCBI	XP_001201249.1	XP_782393.1
Ciona intestinalis	JGI	237751	-
Branchiostoma floridae	JGI	276559	288618
Danio rerio	EnsEMBL	-	ENSDARP00000069804
Mus musculus	EnsEMBL	-	ENSMUSP00000020886
Homo sapiens	EnsEMBL	-	ENSP00000222553

Supplementary Table S2: EST sequences of NAMPT and PNC.

Species		ID	Dev. Stage	Source
B. floridae	PNC	bflv011j05	36 hr larvae	http://amphioxus.icob.sinica.edu.tw/
	NAMPT	bfga038e11	gastrula, neurula	http://amphioxus.icob.sinica.edu.tw/
S. purpuratus	PNC	WHL22.88771.0	embryos, larvae (0-72 hr)	http://www.spbase.org/SpBase/rnaseq/bin/query.html
	NAMPT	WHL22.521505.0	embryos, larvae (0-72 hr)	http://www.spbase.org/SpBase/rnaseq/bin/query.html
N. vectensis	PNC	Nve.21423	unfertilized eggs to primary polyps; whole embryo	http://www.ncbi.nlm.nih.gov/unigene

Supplementary Table S3: MATLAB input data. For each pair of species, mean divergence times are shown in million years as well as protein distances calculated from NAMPT and PNC amino acid alignments. . Hs, *Homo sapiens*; Mm, *Mus musculus*; Dr, *Danio rerio*; Bf, *Branchiostoma floridae*; Ci, *Ciona intestinalis*; Sp, *Strongylocentrotus purpuratus*; Ct, *Capitella teleta*; Hr, *Helobdella robusta*; Lg, *Lottia gigantea*; Dm, *Drosophila melanogaster*; Ce, *Caenorhabditis elegans*; Nv, *Nematostella vectensis*; Ta, *Trichoplax adhaerens*; Sc, *Saccharomyces cerevisiae*; Mb, *Monosiga brevicollis*.

Species pair	Divergence Time (MY)	Protein distance	
		NAMPT	PNC
HsMm	93.9	0.035	-100
HsDr	413.7	0.119	-100
HsBf	713.2	0.483	-100
HsCi	732.8	-100	-100
HsSp	742.9	0.509	-100
HsCt	777.8	0.532	-100
HsHr	777.8	0.589	-100
HsLg	777.8	0.564	-100
HsDm	777.8	-100	-100
HsCe	960.3	-100	-100
HsNv	891.8	0.505	-100
HsTa	940	0.548	-100
HsSc	1232.4	-100	-100
HsMb	857.8	0.576	-100
MmDr	400.1	0.114	-100
MmBf	713.2	0.498	-100
MmCi	722.5	-100	-100
MmSp	742.9	0.509	-100
MmCt	782.7	0.525	-100
MmHr	782.7	0.58	-100
MmLg	782.7	0.556	-100
MmDm	782.7	-100	-100
MmCe	937.5	-100	-100
MmNv	891.8	0.502	-100
MmTa	940	0.56	-100

MmSc	1215.8	-100	-100
MmMb	857.8	0.576	-100
DrBf	713.2	0.494	-100
DrCi	722.5	-100	-100
DrSp	742.9	0.505	-100
DrCt	782.7	0.513	-100
DrHr	782.7	0.597	-100
DrLg	782.7	0.552	-100
DrDm	782.7	-100	-100
DrCe	937.5	-100	-100
DrNv	891.8	0.49	-100
DrTa	940	0.544	-100
DrSc	1215.8	-100	-100
DrMb	857.8	0.556	-100
BfCi	710.5	-100	0.618
BfSp	742.9	0.341	0.607
BfCt	777.8	0.374	0.513
BfHr	777.8	0.436	0.855
BfLg	777.8	0.371	0.944
BfDm	777.8	-100	0.664
BfCe	960.3	-100	0.841
BfNv	891.8	0.325	0.533
BfTa	940	0.377	-100
BfSc	1232.4	-100	0.898
BfMb	857.8	0.367	0.749
CiSp	742.9	-100	0.687
CiCt	782.7	-100	0.749
CiHr	782.7	-100	0.898
CiLg	782.7	-100	0.991
CiDm	782.7	-100	0.749
CiCe	937.5	-100	0.898
CiNv	891.8	-100	0.629
CiTa	940	-100	-100
CiSc	1215.8	-100	0.869
CiMb	857.8	-100	0.749
SpCt	782.7	0.367	0.736
SpHr	782.7	0.494	0.787
SpLg	782.7	0.394	0.898
SpDm	782.7	-100	0.787
SpCe	937.5	-100	0.944
SpNv	891.8	0.397	0.736
SpTa	940	0.443	-100
SpSc	1215.8	-100	1.007
SpMb	857.8	0.45	0.724

CtHr	506	0.443	0.913
CtLg	560.2	0.335	0.944
CtDm	594.8	-100	0.585
CtCe	937.5	-100	0.913
CtNv	891.8	0.312	0.675
CtTa	940	0.367	-100
CtSc	1215.8	-100	1.058
CtMb	857.8	0.415	0.787
HrLg	777.8	0.454	0.841
HrDm	777.8	-100	1.007
HrCe	960.3	-100	1.007
HrNv	891.8	0.502	0.828
HrTa	940	0.505	-100
HrSc	1232.4	-100	1.04
HrMb	857.8	0.505	0.814
LgDm	624	-100	1.075
LgCe	937.5	-100	1.007
LgNv	891.8	0.364	0.855
LgTa	940	0.415	-100
LgSc	1215.8	-100	1.04
LgMb	857.8	0.461	0.991
DmCe	937.5	-100	0.991
DmNv	891.8	-100	0.675
DmTa	940	-100	-100
DmSc	1215.8	-100	1.007
DmMb	857.8	-100	0.652
CeNv	891.8	-100	0.801
CeTa	940	-100	-100
CeSc	1215.8	-100	1.075
CeMb	857.8	-100	0.959
NvTa	940	0.377	-100
NvSc	1215.8	-100	0.869
NvMb	857.8	0.408	0.749
TaSc	1215.8	-100	-100
TaMb	857.8	0.454	-100
ScMb	998.5	-100	1.093

Supplementary Table S4: PDBs used as reference and grid parameters for the docking calculations.

Molecule	Grid Center (mean of C2-C5 atoms distance ligand)	Grid Size (Å)	PDB id	References
NAMPT	X=10.54 Y=-9.25 Z=38.78	30x25x40	2E5D (*H. sapiens*)	(Takahashi et al. 2010)
PNC	X=25.127 Y=−0.983 Z=21.264	35x35x40	3R2J (*Leishmania)*	(Gazanion et al. 2011)

Supplementary Table S5: Residue positions for each species that correspond to the reference residues (NAMPT PDB id 2E5D and PNC PDB id 3R2J) of the alignment used to model the proteins.

Reference molecule	*B. floridae*	*C. teleta*	*N. vectensis*	*S. purpuratus*
NAMPT 2E5D: Active site				
Asp 16 (chain A)	Asp 9 (chain A)	Asp 16 (chain A)	Asp 15 (chain A)	Asp 17 (chain A)
Tyr 18 (chain A)	Tyr 11 (chain A)	Tyr 18 (chain A)	Tyr 17 (chain A)	Tyr 19 (chain A)
Phe 193 (chain B)	Phe 176 (chain B)	Phe 183 (chain B)	Phe 182 (chain B)	Phe 184 (chain B)
Arg 196 (chain B)	Arg 179 (chain B)	Arg 186 (chain B)	Arg 185 (chain B)	Arg 187 (chain B)
Asp 219 (chain B)	Asp 202 (chain B)	Asp 209 (chain B)	Asp 208 (chain B)	Asp 210 (chain B)
Arg 311 (chain B)	Arg 293 (chain B)	Arg 300 (chain B)	Arg 298 (chain B)	Arg 301 (chain B)
Active site (3R2J)				
Leu-20	Leu27	Met27	Leu16	Leu27
Val-22	Val29	Ile29	Leu18	Leu29
Trp-91	Trp110	Trp110	Trp99	Trp106
Tyr-131	Tyr147	Tyr147	Tyr137	Tyr143
Ala-163	Ala179	Ala179	Ala169	Ala175
Tyr-166	Val182	Val182	Tyr172	Val178
Ile-192	Val208	Val208	Ile198	Gln204
Catalytic (3R2J)				
Asp-8	Asp17	Asp17	Asp6	Asp17
Lys-122	Lys138	Lys138	Lys128	Lys134
Cys-167	Cys183	Cys183	Cys173	Cys179

Supplementary Table S6: Oligonucleotide sequences used for amplification of NAMPT and PNC from B. floridae, S. purpuratus, C. teleta and N. vectensis.

Species	Gene	Forward primer	Reverse primer	AT
		(5'-3')	(5'-3')	(ºC)
B. floridae	PNC	CTCATAGTGGTGGACATGCA	CAGAATGCCGAGTAACTGTC	58
	NAMPT	CATCACGGACTCCTACAAGG	TTGCGATCGTGTCTGTCCC	59
S. purpuratus	PNC	CCTCATAGCAGTAGATGTAC	GCAGAATAGCTGTCGACATG	52
	NAMPT	GACGGTTCTTACAAGGTCAC	CCTGCTATGGTGTCTGTTCC	59
C. teleta	PNC	ATTGGTAATCGTGGACGTGC	AGAACGCTGAGTAGCTGTC	62
	NAMPT	CCTATAAGGTGACCCATCAC	CATAGTACTTGCGAGCCGTC	62
N. vectensis	PNC	CCTGATAGTTGTTGATGTACA	AGAAGGCAGAGTAGCTGTC	60
	NAMPT	CGGATTCGTACAAGGTCTCC	CAGCTATCGTGTCTGTGCC	62

References

1. Revollo JR, Grimm AA, Imai S (2007) The regulation of nicotinamide adenine dinucleotide biosynthesis by Nampt/PBEF/visfatin in mammals. Curr Opin Gastroenterol 23: 164-170.
2. Belenky P, Christensen KC, Gazzaniga F, Pletnev AA, Brenner C (2009) Nicotinamide riboside and nicotinic acid riboside salvage in fungi and mammals. Quantitative basis for Urh1 and purine nucleoside phosphorylase function in NAD+ metabolism. J Biol Chem 284: 158-164.
3. Bogan KL, Brenner C (2008) Nicotinic acid, nicotinamide, and nicotinamide riboside: a molecular evaluation of NAD+ precursor vitamins in human nutrition. Annu Rev Nutr 28: 115-130.
4. Belenky P, Bogan KL, Brenner C (2007) NAD+ metabolism in health and disease. Trends Biochem Sci 32: 12-19.
5. Hara N, Yamada K, Shibata T, Osago H, Hashimoto T, et al. (2007) Elevation of cellular NAD levels by nicotinic acid and involvement of nicotinic acid phosphoribosyltransferase in human cells. J Biol Chem 282: 24574-24582.
6. Magni G, Amici A, Emanuelli M, Orsomando G, Raffaelli N, et al. (2004) Enzymology of NAD+ homeostasis in man. Cell Mol Life Sci 61: 19-34.
7. Gazanion E, Garcia D, Silvestre R, Gerard C, Guichou JF, et al. (2011) The Leishmania nicotinamidase is essential for NAD+ production and parasite proliferation. Mol Microbiol 82: 21-38.
8. Jewett MW, Jain S, Linowski AK, Sarkar A, Rosa PA (2011) Molecular characterization of the Borrelia burgdorferi *in vivo*-essential protein PncA. Microbiology 157: 2831-2840.
9. Li YF, Bao WG (2007) Why do some yeast species require niacin for growth? Different modes of NAD synthesis. FEMS Yeast Res 7: 657-664.
10. Martin PR, Shea RJ, Mulks MH (2001) Identification of a plasmid-encoded gene from Haemophilus ducreyi which confers NAD independence. J Bacteriol 183: 1168-1174.
11. Anderson RM, Bitterman KJ, Wood JG, Medvedik O, Sinclair DA (2003) Nicotinamide and PNC1 govern lifespan extension by calorie restriction in Saccharomyces cerevisiae. Nature 423: 181-185.
12. Gallo CM, Smith DL, Jr., Smith JS (2004) Nicotinamide clearance by Pnc1 directly regulates Sir2-mediated silencing and longevity. Mol Cell Biol 24: 1301-1312.
13. Revollo JR, Grimm AA, Imai S (2004) The NAD biosynthesis pathway mediated by nicotinamide phosphoribosyltransferase regulates Sir2 activity in mammalian cells. J Biol Chem 279: 50754-50763.
14. Galli M, Van Gool F, Rongvaux A, Andris F, Leo O (2010) The nicotinamide phosphoribosyltransferase: a molecular link between metabolism, inflammation, and cancer. Cancer Res 70: 8-11.
15. Mesko B, Poliska S, Szegedi A, Szekanecz Z, Palatka K, et al. (2010) Peripheral blood gene expression patterns discriminate among chronic inflammatory diseases and healthy controls and identify novel targets. BMC Med Genomics 3: 15.
16. Balan V, Miller GS, Kaplun L, Balan K, Chong ZZ, et al. (2008) Life span extension and neuronal cell protection by Drosophila nicotinamidase. J Biol Chem 283: 27810-27819.
17. Burnett C, Valentini S, Cabreiro F, Goss M, Somogyvari M, et al. (2011) Absence of effects of Sir2 overexpression on lifespan in C. elegans and Drosophila. Nature 477: 482-485.
18. Silva RM, Duarte IC, Paredes JA, Lima-Costa T, Perrot M, et al. (2009) The yeast PNC1 longevity gene is up-regulated by mRNA mistranslation. PLoS One 4: e5212.
19. van der Horst A, Schavemaker JM, Pellis-van Berkel W, Burgering BM (2007) The Caenorhabditis elegans nicotinamidase PNC-1 enhances survival. Mech Ageing Dev 128: 346-349.
20. French JB, Cen Y, Vrablik TL, Xu P, Allen E, et al. (2010) Characterization of nicotinamidases: steady state kinetic parameters, classwide inhibition by nicotinaldehydes, and catalytic mechanism. Biochemistry 49: 10421-10439.

21. Zhang JL, Zheng QC, Li ZQ, Zhang HX (2012) Molecular Dynamics Simulations Suggest Ligand's Binding to Nicotinamidase/Pyrazinamidase. PLoS One 7: e39546.
22. Domergue R, Castano I, De Las Penas A, Zupancic M, Lockatell V, et al. (2005) Nicotinic acid limitation regulates silencing of Candida adhesins during UTI. Science 308: 866-870.
23. Garten A, Petzold S, Korner A, Imai SI, Kiess W (2009) Nampt: linking NAD biology, metabolism and cancer. Trends Endocrinol Metab 20: 8.
24. Khan JA, Tao X, Tong L (2006) Molecular basis for the inhibition of human NMPRTase, a novel target for anticancer agents. Nat Struct Mol Biol 13: 582-588.
25. Kim MK, Lee JH, Kim H, Park SJ, Kim SH, et al. (2006) Crystal structure of visfatin/pre-B cell colony-enhancing factor 1/nicotinamide phosphoribosyltransferase, free and in complex with the anti-cancer agent FK-866. J Mol Biol 362: 66-77.
26. Olesen UH, Petersen JG, Garten A, Kiess W, Yoshino J, et al. (2010) Target enzyme mutations are the molecular basis for resistance towards pharmacological inhibition of nicotinamide phosphoribosyltransferase. BMC Cancer 10: 677.
27. Zhang LY, Liu LY, Qie LL, Ling KN, Xu LH, et al. (2012) Anti-proliferation effect of APO866 on C6 glioblastoma cells by inhibiting nicotinamide phosphoribosyltransferase. Eur J Pharmacol 674: 163-170.
28. Ma B, Pan SJ, Zupancic ML, Cormack BP (2007) Assimilation of NAD(+) precursors in Candida glabrata. Mol Microbiol 66: 14-25.
29. Seiner DR, Hegde SS, Blanchard JS (2010) Kinetics and inhibition of nicotinamidase from Mycobacterium tuberculosis. Biochemistry 49: 9613-9619.
30. Sorci L, Blaby I, De Ingeniis J, Gerdes S, Raffaelli N, et al. (2010) Genomics-driven reconstruction of acinetobacter NAD metabolism: insights for antibacterial target selection. J Biol Chem 285: 39490-39499.
31. Lin H, Kwan AL, Dutcher SK (2010) Synthesizing and salvaging NAD: lessons learned from Chlamydomonas reinhardtii. PLoS Genet 6.
32. Gazzaniga F, Stebbins R, Chang SZ, McPeek MA, Brenner C (2009) Microbial NAD metabolism: lessons from comparative genomics. Microbiol Mol Biol Rev 73: 529-541, Table of Contents.
33. Gossmann TI, Ziegler M, Puntervoll P, de Figueiredo LF, Schuster S, et al. (2012) NAD(+) biosynthesis and salvage - a phylogenetic perspective. FEBS J.
34. Miller ES, Heidelberg JF, Eisen JA, Nelson WC, Durkin AS, et al. (2003) Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. J Bacteriol 185: 5220-5233.
35. Burgos ES, Ho MC, Almo SC, Schramm VL (2009) A phosphoenzyme mimic, overlapping catalytic sites and reaction coordinate motion for human NAMPT. Proc Natl Acad Sci U S A 106: 13748-13753.
36. Takahashi R, Nakamura S, Nakazawa T, Minoura K, Yoshida T, et al. (2010) Structure and reaction mechanism of human nicotinamide phosphoribosyltransferase. J Biochem 147: 95-107.
37. Wang T, Zhang X, Bheda P, Revollo JR, Imai S, et al. (2006) Structure of Nampt/PBEF/visfatin, a mammalian NAD+ biosynthetic enzyme. Nat Struct Mol Biol 13: 661-662.
38. Du X, Wang W, Kim R, Yakota H, Nguyen H, et al. (2001) Crystal structure and mechanism of catalysis of a pyrazinamidase from Pyrococcus horikoshii. Biochemistry 40: 14166-14172.
39. Hu G, Taylor AB, McAlister-Henn L, Hart PJ (2007) Crystal structure of the yeast nicotinamidase Pnc1p. Arch Biochem Biophys 461: 66-75.
40. Fyfe PK, Rao VA, Zemla A, Cameron S, Hunter WN (2009) Specificity and mechanism of Acinetobacter baumanii nicotinamidase: implications for activation of the front-line tuberculosis drug pyrazinamide. Angew Chem Int Ed Engl 48: 9176-9179.

41. Petrella S, Gelus-Zietal N, Maudry A, Laurans C, Boudjelloul R, et al. (2011) Crystal structure of the pyrazinamidase of Mycobacterium tuberculosis: insights into natural and acquired resistance to pyrazinamide. PLoS One 6: e15785.
42. French JB, Cen Y, Sauve AA, Ealick SE (2010) High-resolution crystal structures of Streptococcus pneumoniae nicotinamidase with trapped intermediates provide insights into the catalytic mechanism and inhibition by aldehydes. Biochemistry 49: 8803-8812.
43. Irimia M, Tena JJ, Alexis MS, Fernandez-Minan A, Maeso I, et al. (2012) Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. Genome Research.
44. Louis A, Roest Crollius H, Robinson-Rechavi M (2012) How much does the amphioxus genome represent the ancestor of chordates? Brief Funct Genomics 11: 89-95.
45. Azevedo L, Carneiro J, van Asch B, Moleirinho A, Pereira F, et al. (2009) Epistatic interactions modulate the evolution of mammalian mitochondrial respiratory complex components. BMC Genomics 10: 266.
46. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.
47. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731-2739.
48. Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. Edited in Evolving Genes and Proteins by V Bryson and HJ Vogel Academic Press, New York: 97-166.
49. Drummond A, Ashton B, Buxton S, Cheung M, Cooper A, et al. (2011) Geneious. v5.5 ed: http://www.geneious.com.
50. Hedges SB, Dudley J, Kumar S (2006) TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22: 2971-2972.
51. Wang Z, Ding GH, Yu ZH, Liu L, Li YX (2009) CHSMiner: a GUI tool to identify chromosomal homologous segments. Algorithms for Molecular Biology 4.
52. Schrödinger LLC (2009) Prime, version 2.1, New York, NY.
53. Baldi P, Brunak S, Frasconi P, Soda G, Pollastri G (1999) Exploiting the past and the future in protein secondary structure prediction. Bioinformatics 15: 937-946.
54. Pollastri G, Przybylski D, Rost B, Baldi P (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 47: 228-235.
55. Schrödinger LLC (2011) LigPrep, version 2.5, New York, NY.
56. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, et al. (2009) AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem 30: 2785-2791.
57. Sanner MF (1999) Python: a programming language for software integration and development. J Mol Graph Model 17: 57-61.
58. Pohorille A, Jarzynski C, Chipot C (2010) Good practices in free-energy calculations. J Phys Chem B 114: 10235-10253.
59. Shirts MR, Pitera JW, Swope WC, Pande VS (2003) Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins Journal of Chemical Physics 119: 5740-5761.
60. Wallace AC, Laskowski RA, Thornton JM (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng 8: 127-134.
61. McDonald IK, Thornton JM (1994) Satisfying hydrogen bonding potential in proteins. J Mol Biol 238: 777-793.
62. Schrödinger LLC (2010) The PyMOL Molecular Graphics System, Version 1.3r1 (http://pymol.org/).
63. Ocean Genome Legacy, Ocean Genome Resource database. Published on the Web at: www.oglf.org/Catalog.htm; accessed September 2011.

64. Yu JK, Wang MC, Shin IT, Kohara Y, Holland LZ, et al. (2008) A cDNA resource for the cephalochordate amphioxus Branchiostoma floridae. Dev Genes Evol 218: 723-727.
65. Tu Q, Cameron RA, Worley KC, Gibbs RA, Davidson EH (2012) Gene structure in the sea urchin Strongylocentrotus purpuratus based on transcriptome analysis. Genome Res 22: 2079-2087.
66. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2007) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 35: D5-12.

9. Publication III: Mitochondrial DNA deletions are associated with non-B DNA conformations

Mitochondrial DNA deletions are associated with non-B DNA conformations

Joana Damas,[1] João Carneiro,[1,2] Joana Gonçalves,[1] James B Stewart,[3] David C Samuels,[4] António Amorim,[1,2] Filipe Pereira[1*]

[1] Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

[2] Faculty of Sciences, University of Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

[3] Max Planck Institute for Biology of Ageing, Cologne, Germany

[4] Center for Human Genetic Research, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America

*To whom correspondence should be addressed:

Filipe Pereira

IPATIMUP. Rua Dr. Roberto Frias s/n 4200-465 Porto, PORTUGAL

Phone: +351 22 5570700 / Fax: +351 22 5570799

E-mails: fpereirapt@gmail.com

Keywords: mitochondrial DNA, deletions, breakpoints, non-B DNA, hairpins, cloverleaf structures, mitochondrial disease

Running title: The role of non-B DNA in mtDNA deletions

Nucleic Acids Res. 2012 September; 40(16): 7606–7621.

Published online 2012 May 31. doi: 10.1093/nar/gks500

Abstract

Mitochondrial DNA (mtDNA) deletions are a primary cause of mitochondrial disease and are believed to contribute to the aging process and to various neurodegenerative diseases. Despite strong observational and experimental evidence, the molecular basis of the deletion process remains obscure. In this study, we test the hypothesis that the primary cause of mtDNA vulnerability to breakage resides in the formation of non-B DNA conformations, namely hairpin, cruciform and cloverleaf-like elements. Using the largest database of human mtDNA deletions built thus far (753 different cases), we show that site-specific breakage hotspots exist in the mtDNA. Furthermore, we discover that the most frequent deletion breakpoints occur within or near predicted structures, a result that is supported by data from transgenic mice with mitochondrial disease. There is also a significant association between the folding energy of an mtDNA region and the number of breakpoints that it harbours. In particular, two clusters of hairpins (near the D-loop 3' terminus and the L-strand origin of replication) are hotspots for mtDNA breakage. Consistent with our hypothesis, the highest number of 5' and 3' breakpoints per base is found in the highly structured tRNA genes. Overall, the data presented in this study suggest that non-B DNA conformations are a key element of the mtDNA deletion process.

Introduction

It is undeniable that the complementary strands (named the L- and H-strands) of human mitochondrial DNA (mtDNA) are primarily organised into the canonical right-handed double-helical structure of B-form DNA. Unfortunately, our knowledge of the higher-order topology of mtDNA and its interaction with the mitochondrial environment remains quite limited. It is plausible that many deformations to the canonical B-form of DNA occur in the mitochondrial genome and have important biological consequences, as has been unequivocally shown in many other genetic systems (1-3). In recent years, several studies have demonstrated that non-B DNA structures (often called noncanonical, unusual, alternative or secondary DNA structures) (4-6) occur, at least transiently, in the mitochondrial genome. For instance, a stem-loop structure is required to activate the initiation of DNA replication in the L-strand origin of replication (O_L) (7;8). DNA bending in the L-strand promoter (LSP) induced by the mitochondrial transcription factor A (TFAM) is necessary for transcription initiation (9;10), whereas DNA unwinding and base eversion at the *tRNA-Leu(UUA/G)* gene by the mitochondrial transcription termination factor 1 (MTERF1) is critical for transcription termination (11).

The formation of most non-B DNA structures is favoured by the local unwinding of the DNA double helix, which is associated with negative supercoiling (4;12;13). As is most DNA *in vivo*, mtDNA is believed to be predominantly negatively supercoiled (i.e., the torsional tension diminishes the DNA helicity and facilitates strand separation) and is subject to dynamic processes that constantly alter the canonical conformation of the double helix. Unlike nuclear DNA, mtDNA is continuously replicated and transcribed during the entire cell cycle and, according to the available models, a large portion of the mtDNA is single-stranded for a significant period of time during such processes (14-16). This phenomenon provides an opportunity for structures with intra-strand base pairing to form and to persist for a relatively long period. In addition, several proteins are continuously tracking through the mtDNA (e.g., the movement of an RNA polymerase during transcription), resulting in a redistribution of the local supercoiling characteristics, which can be counterbalanced by the action of topoisomerases or the formation of non-B DNA structures (17;18). The conformational flexibility of the mtDNA is also affected by the packaging factor TFAM, which induces negative supercoiling upon binding to mtDNA (10;19). Thus, the binding of proteins may restrict the transmission of the superhelical tension throughout the DNA, although this possibility remains to be determined for the mitochondrial genome. Similarly, stable and partially hybridised RNA molecules (R-loops) were found to be associated with the mammalian mtDNA and are thought to maintain the genome in a more open conformation (20).

The versatile nature of the mitochondrial genome is also evident in the multiple forms of gene organisation and structural diversity resulting from numerous genomic rearrangements that occur through insertions, deletions, duplications, inversions or translocations of DNA segments (21). Among the different types of rearrangements, the loss of a section of the mitochondrial genome (an mtDNA deletion) has attracted the attention of researchers. The reason for such concern is that mtDNA deletions are associated with the multifactorial aging process and with a variety of progressive disorders that cause substantial disability and can lead to premature death (16;22;23). The loss of mtDNA-encoded proteins and/or tRNA genes required for protein synthesis results in mitochondrial dysfunction and cell death due to their crucial role in energy metabolism. In order to cause mitochondrial dysfunction, a deleted species of mtDNA must accumulate above a critical threshold that varies from tissue to tissue based on energy requirements. The expansion of a deleted mtDNA molecule to the detriment of other variants might occur due to the genetic bottleneck for mtDNA transmission in the germline and the unequal portioning of molecules in daughter cells (mitotic segregation). The level of different types of mtDNA molecules within a cell (heteroplasmy) might change randomly (intracellular drift) and independently of the cell cycle by a process of relaxed replication (24-27).

However, the exact mechanism(s) underlying the formation of mtDNA deletions remain elusive. What is undeniable is that most mtDNA deletions occur in the major arc of the mtDNA, between the two proposed origins of replication (O_H and O_L), and that they present sequence homologies at the boundaries of their breakpoints (28-30). In addition, several authors have noticed that mtDNA deletion breakpoints are often located in regions with the potential to adopt non-B DNA conformations, raising the possibility of a mechanistic role of alternative DNA structures in the generation of deletions (30-38). In this study, we describe a comprehensive survey of non-B DNA conformations across the complete human mitochondrial genome and provide multiple lines of evidence for their association with mtDNA deletions. Our study was prompted by the clear recognition that certain DNA sequences adopt structural configurations that are more prone to breakage (39-43) and the emerging understanding that non-B DNA structures are inherently hypermutable (1-3;44).

Material and Methods

Mitochondrial DNA deletion breakpoints

We started by collecting information about all of the available mtDNA deletions (5' and 3' breakpoints) from the MITOMAP (http://www.mitomap.org) and MitoTool (45) databases and from 83 peer reviewed papers published from 1989 to 2010 (Supplemental Fig. S1). The 929 different deletions that were initially retrieved have been identified in a) patients with an mtDNA deletion syndrome (chronic progressive external ophthalmoplegia, Kearns-Sayre syndrome or Pearson syndrome) or with a complex multi-system disorder that did not fit into any of the preceding categories; b) patients with autosomal disorders of mtDNA maintenance or mitochondrial nucleotide metabolism and c) post-mitotic tissues as part of normal aging. We decided to combine deletions from different sources because there is no evidence so far that different molecular mechanisms cause deletions in the different clinical scenarios that would justify a separate analysis. It has been suggested that a similar mechanism generates mtDNA deletions in all clinical situations (29).

Each deletion was defined by a unique combination of two breakpoints and was only included once in our database. This procedure was used to avoid the ascertainment bias that is caused by the regular use of methods that only identify a restricted group of deletions, which would lead to an overrepresentation of some deletions in our database if frequency values were considered. Moreover, we attempted to minimise the noise inherent to the presence of the same deletion in different databases or publications (sometimes even with a different nomenclature) that would cause an artificial duplication of data.

If different deletions shared the same breakpoint at one end, then the shared breakpoint was counted once for each deletion in most analyses. The deletion breakpoints were always numbered according to the conventional L-strand positions of the revised Cambridge reference mtDNA sequence (rCRS; NC_012920). We always considered, in this study, that 'breakpoints' are the mtDNA positions that are retained in the deleted mtDNA sequence and that flank the deleted region. In other words, 5' breakpoints are upstream of the 5' break and 3' breakpoints are downstream of the 3' break, considering the L-strand numbering.

In several cases, we observed that mtDNA deletions are described in the literature by an interval of values as breakpoint positions. This type of nomenclature is used because of the existence of equal sequence motifs in the breakpoint areas (for example, 8,016-8,019:15,516-15,519), which renders the precise identification of the break sites impossible. In such situations, we have retained the smallest number for each breakpoint in the interval (in the previous example, 8,016:15,516). With this correction, a few deletions were found to be repeated in our original database and were removed, leaving a total of 788 deletions. We

then generated the 788 deleted mtDNA sequences by removing the region between the 5' and 3' breakpoints in the rCRS, retrieved from the NCBI Entrez Nucleotide database (http://www.ncbi.nlm.nih.gov). Each deleted sequence was aligned with the full-length rCRS using python scripts (Python v.2.6; www.python.org/) from the 3rd party application 'Muscle' (46) available on the PyCogent v1.5 package (47). The existence of equal sequence motifs in breakpoint areas implies that different sequence alignments are possible for the same deletion: certain bases might be equally aligned upstream or downstream of the deletion area, in the equal sequence motifs (Supplemental Fig. S2). Therefore, we have corrected the limits of the deletion when necessary by keeping all possible matches at the 5' breakpoint. This adjustment revealed that several deletions with different reported breakpoints were in fact equal. As a result, 753 unique deletions remained and were used in all of the analyses.

In addition, we collected all of the available information on the mtDNA breakpoints of transgenic mice expressing an altered Twinkle mtDNA helicase (48). The breaking sites were plotted on the *Mus musculus* mtDNA reference sequence (NC_005089), positions 15,150 to 15,469.

Prediction of non-B DNA conformations

In this work, we use four expressions to describe the predicted structural alterations to the orthodox right-handed Watson-Crick B-form of mtDNA: hairpin, cruciform, cloverleaf-like elements and secondary structures. All of these terms refer to deviations from the conventional B-form of DNA, collectively known as non-B DNA conformations. A hairpin or stem-loop structure is a section of single-stranded DNA that folds back on itself to form a paired double helix that ends in an unpaired loop. The cruciform structure consists of a pair of hairpin structures in complementary DNA strands forming a four-way junction with a cross-shaped configuration. The cloverleaf structure is a single-strand DNA arrangement with four stems and three terminal loops, usually used to describe the secondary structure of tRNA molecules. All other types of structural arrangements that do not belong to the preceding categories are designated here as 'secondary structures'. The terms are used regardless of the length and folding energy of the stem and loop regions. It should be taken into consideration that other types of non-B DNA elements exist and were not addressed here (e.g., triplexes, G4-tetrad, slipped structures, left-handed Z-DNA, bent DNA and sticky DNA) (2).

The folding prediction for single-stranded DNA was performed with the hybrid-ss-min core programme of the UNAFold software package (49). Python scripts were written to run automatic executables from UNAFold. The prediction is based on free energy minimisation using nearest neighbour thermodynamic rules and dynamic programming algorithms (50). The folding of the single-stranded DNA was carried out using the default parameters,

including predictions at a temperature of 37°C, sodium concentration of 1 M and magnesium concentration of 0 M. The hybrid-ss-min programme predicts the thermodynamically most stable secondary structure that a single-stranded DNA segment can form and calculates the variation in the free energy of the folding (ΔG, expressed here in kcal/mol). The magnitudes of the ΔG values indicate the relative stabilities of the structures formed by each segment: the greater the variation in the ΔG value (more negative value), the more likely it is that a stable secondary structure will form. We always used the predicted lowest free energy structure (the structure with the most negative ΔG value) in the various analyses, although other suboptimal structures were sometimes predicted by the programme. The graphical representations of DNA secondary structures were obtained from the sir-graph programme, which belongs to the mfold-util software v4.6 (51). The circular maps of the human mtDNA were produced using the Circos software, version 0.52 (52).

Statistical analyses

The descriptive statistics for the different datasets, the Student's *t*-test (independent samples with separate variance estimates) and the Fisher´s exact test for contingency tables were obtained with the STATISTICA v7 software (StatSoft, Inc., Tulsa, OK). All reported *p*-values are two-sided and a significance level of 0.05 was used.

Results

Deletion breakpoints are not randomly distributed throughout the mitochondrial genome

We started by collecting data from all of the available mtDNA deletions in public databases and peer-reviewed publications (Supplemental Fig. S1) that have been identified in pathological and non-pathological situations. The 753 unique mtDNA deletions that fulfilled our selection criteria (Supplemental Fig. S3; see Methods) are defined by 620 and 497 different 5' and 3' breakpoints, respectively (1,117 different breakpoint in total). The number of breakpoints is lower than 1,506 (i.e., twice the total number of 753 deletions) because different deletions sometimes share the same breakpoint at one end. The distributions of the 5' and 3' breakpoints across the mtDNA are clearly different from each other (Fig. 1; Supplemental Figs. S4, S5). This difference has been previously noted in smaller deletion datasets (28;30;53).

The mean value of the distribution of the 5' breakpoints is position 7,658 with a standard deviation of 2,296 (Supplemental Fig. S4). The mode is position 7,402 with a total of eight breakpoints, although it is not clearly distinct from the values at other positions (e.g., positions 5,787 and 8,032, with seven and six breakpoints, respectively). The histogram of the distribution of 5' breakpoints suggests a multimodal distribution with major peaks around and within *COX2* and in the WANCY cluster of tRNA genes (Fig. 1; Supplemental Fig. S4).

The mean value of the distribution of the 3' breakpoints is position 14,503, with a standard deviation of 2,185 (Fig. 1; Supplemental Fig. S5). There is a clear mode in the distribution: position 16,071 has noticeably the highest number of 3' breakpoints, with 41 out of 753 total breakpoints (5.44%). The flanking sites of the 16,071 hotspot (positions 16,065 to 16,080) harbour 25.09% of all of the 3' breakpoints (189 out of 753 breakpoints) (Supplemental Fig. S6). The remaining deletions are mainly found in the *ND5* and *CYTB* genes, although there is a sudden decrease in the number of 3' breakpoints between them, in the region of the *ND6* and *tRNA-Glu* genes (Fig. 1; Supplemental Fig. S5). Overall, the mtDNA deletions are not randomly distributed, and their breakpoints do not follow a normal distribution around any specific mtDNA position (Fig. 1).

A completely different pattern would be expected if the deletions were random. As a simple way to estimate the frequency of deletions per site, we generated a dataset of 20,000 random deletions with no restrictions. The distribution of mtDNA breakage hotspots in the real data contrasts with the distribution observed in simulated deletions. The most common breakpoints have frequency values [e.g., mtDNA positions 16,071 ($5.4x10^{-2}$), 7,402 ($1.1x10^{-2}$), 5,787 ($9.3x10^{-3}$) and 15,435 ($8.0x10^{-3}$)] that are considerably higher than those estimated for 20,000 random deletions, where the highest breakpoint frequency at any site is $4.0x10^{-4}$ (8

occurrences out of 20,000 breakpoints) (Supplemental Fig. S7). The most frequent real and simulated breakpoints have a significantly different proportion in our datasets (p-value < 1.00×10^{-4}; Student's t-test): 8/753 = 0.011 and 41/753 = 0.054 for 5' and 3' breakpoints, respectively; 8/20,000 = 0.0004 for random deletions.

The most frequent deletion breakpoints occur within or near predicted hairpins

We started by investigating the locations of the ten main mtDNA breakage hotspots in the predicted secondary structure of the 100-nt windows (L- and H-strands) enclosing each of these breakpoints (selected as window midpoints, i.e., each window extends 50 nt upstream and downstream of the break) (Fig. 2; Supplemental Figs. S7-S9). The free energy of folding of the 100-nt windows varied from -13.19 to -3.05 kcal/mol in the L-strand and from -14.47 to -1.34 kcal/mol in the H-strand. Although all of the 100-nt windows present at least one stem element, four of these regions (3,263, 5,787, 12,300 and 16,071) stand out as being highly structured, with several stem elements and a folding energy lower than -9 kcal/mol in the L- and H-strands (Fig. 2). In most cases, similar hairpins were predicted for both mtDNA strands, which is compatible with the formation of cruciform structures. Cloverleaf structures were predicted for the mtDNA regions of breakpoints 3,263 (L-strand) and 16,071 (L- and H-strands). Similar DNA structures were predicted using 300-nt windows (also with breakpoints as window midpoints), suggesting that short-distance interactions are more stable than long-distance base interactions (data not shown).

The most frequent breakpoints are located in single-stranded regions, such as terminal loops, with the exception of breakpoints at positions 8,387 (L- and H-strand structures) and 7,402 (H-strand structure), which are located in stem regions (Fig. 2; Supplemental Figs. S8, S9). Nevertheless, when considering the 1,117 different breakpoints, we found that several breaks occur between bases that are paired in folded 100-nt sliding windows (Supplemental Fig. S10). The exact location of the 3' breakpoint of the 'common deletion' (13,447) is impossible to ascertain due to the presence of a 13-nt direct repeat at the breakpoint regions. The 100-nt window around this breakpoint folds differently in L- and H-strands, with four and three hairpins, respectively. The breaking of the DNA might occur in a stem or loop element. Other structures predicted for the L- and H-strands also differ in the number and location of hairpins. For instance, the 7,402 and 8,032 breakpoints are located in hairpins elements predicted only for the H-strand (Fig. 2; Supplemental Figs. S8, S9).

To verify whether the association between breakpoints and hairpins was not only a feature of human mitochondrial genome, we analysed the distribution of breakpoints in the mtDNA of transgenic mice with mitochondrial disease (48). The distribution of the deletion breakpoints in the *Mus musculus* mtDNA (position 15,150- 15,469) is biased towards hairpins, with all of the breakpoints occurring in this category of non-B DNA (Fig. 3). The

difference between the proportion of breakpoints located inside (13 breakpoints in a total of 229 nt) and outside (0 breakpoints in a total of 91 nt) hairpins attains statistical significance (*p*-value = 0.023; Fisher's exact test).

Two stable clusters of hairpins are hotspots for mtDNA breakage

The major mtDNA deletion breakpoint (16,071) is located in a 93-nt cloverleaf-like structure (positions 16,028 to 16,120) that we previously identified and named structure A (35). The central hairpin of this structure (16,060 to 16,082) is a hotspot of DNA breakage, with 189 reported 3' breakpoints occurring in this short region (Fig. 4A). The central hairpin has a higher proportion of sites with breakpoints (16 sites out of 23) than the remaining structure (8 sites out of 70; *p*-value < 1.00×10^{-4}; Fisher´s exact test). Of these 189 breakpoints, 144 (19% of all 3' breakpoints) are located on the 8-nt terminal loop (*p*-value < 1.00×10^{-4}; Fisher´s exact test). The 16,071 hotspot is located upstream of the trinucleotide stop point (16,104-16,106) for the premature arrest of the H-strand synthesis (D-loop 3' end), according to the numbering of the mtDNA, or downstream according to the direction of the H-strand synthesis. The 16,071 hotspot is not within the three-stranded D-loop structure (Fig. 4A). It could be hypothesised that the 16,071 3' breakpoint hotspot is overrepresented in our database by being associated with certain particular types of 5' breakpoint in a narrow region of the mitochondrial genome. However, we observed that deletions with a 3' breakpoint in the 16,071 hotspot (16,075 – 16,080) have 5' breakpoints in very different mtDNA regions (Supplemental Fig. S11).

One of the most relevant 5' breakpoint hotspots is located in the WANCY region (Fig. 1, 4B). This region comprises a cluster of five tRNA genes (*tRNA-Trp*, *tRNA-Ala*, *tRNA-Asn*, *tRNA-Cys* and *tRNA-Tyr*) that is located between the *ND2* and *COX1* genes and that includes the O_L. We found that this mtDNA segment (5,512 to 5,903) has a very high folding potential (ΔG = -36.41 kcal/mol for the L-strand and ΔG = -40.49 kcal/mol for the H-strand), comprising several hairpin structures (Fig. 4B; Supplemental Fig. S12). We discovered that all of the 5' ends of the deletions identified in this region (*n* = 27) are located in five of these hairpin elements (Fig. 4B; Supplemental Fig. S12). We observed that the difference in the proportion of breakpoints inside (16 in a total of 295 nt) and outside (0 in a total of 97 nt) hairpins is statistically significant (*p*-value = 0.015; Fisher´s exact test). In particular, 23 of the deletion breakpoints are located in a single stem-loop element predicted for positions 5,772 to 5,803 in the *tRNA-Cys* gene. This element is only 5 bases downstream of the previously identified stem-loop structure that is associated with the origin of L-strand replication (7;8).

Mitochondrial tRNA genes are hotspots for mtDNA breakage

We investigated the distribution of the deletion breakpoints according to the coding/non-coding features of mtDNA. The mtDNA regions with the highest number of 5' breakpoints are the *COX2* and *COX1* genes, with 182 and 149 5' breakpoints, respectively (Supplemental Figs. S13, S14). The control region (the largest non-coding region of mtDNA located between *tRNA-Pro* and *tRNA-Phe* genes) and the *ND5* gene have the highest number of 3' breakpoints (219 breakpoints each). When considering both 5' and 3' breakpoints, the control region is the location where more breaks occur (15.47% of the total breakpoints), followed by the *ND5* (14.74%), *COX2* (12.22%) and *CYTB* (11.82%) genes.

Strikingly, when we take into account the lengths of the different coding or non-coding mtDNA regions, those with the highest number of 5' and 3' breakpoints per base (number of deletions/region length) are the tRNA genes (Fig. 5). *tRNA-Ser(UCN)* and *tRNA-Cys* are the genes with more 5' breakpoints per base (0.406 and 0.348, respectively) (Supplemental Fig. S14). Although they have the highest absolute number of 5' breakpoints, the *COX1* and *COX2* genes only have 0.097 and 0.266 5' breakpoints per base, respectively. The *tRNA-Ser(UCN)* gene has a higher proportion of sites with breakpoints than its adjacent *COX1* gene (p-value < 1.00×10^{-4}; Student's t-test). Inside the minor arc, the *tRNA-Leu(UUA/G)* gene, which encodes the MTERF1 binding site, stands out as having a significantly higher number of 5' breakpoints per base (0.147) than its flanking genes (*RNR2* and *ND1* with 0.013 and 0.023, respectively; p-values < 1.00×10^{-4}; Student's t-test). Similarly, the gene with the highest number of 3' breakpoints per base is *tRNA-Thr* (0.318) (Supplemental Fig. S14). The other regions with the highest number of 3' breakpoints per base are the control region (0.195) and the *CYTB* gene (0.152). Nevertheless, the proportion of sites with breakpoints is significantly higher in the *tRNA-Thr* gene than in the adjacent *CYTB* (p-value = 6.00×10^{-4}; Student's t-test; Fig. 5).

Overall, the tRNA gene sequences fold into structures quite different from the common tRNA cloverleaf structure with four stems and three loops (Supplemental Figs. S15, S16). In several cases, the predicted structures lack complete domains of the cloverleaf structure such as the acceptor arm (e.g., *tRNA-Ala*, *tRNA-Arg* or *tRNA-Asn*). The differences between structures are explained by the different folding proprieties of DNA and RNA molecules. We also found that the variation in folding energies among tRNA genes is not sufficient to explain the difference in the frequency of breakpoints (Supplemental Figs. S17). The mtDNA regions with more breakpoints per base (5' and 3') are the *tRNA-Ser(UCN)* (0.406), *tRNA-Cys* (0.348) and *tRNA-Thr* (0.318) genes (Fig. 5). These three breakpoint-prone tRNA genes fold with the formation of at least two stem elements. The *tRNA-Thr* gene is one of the few cases that are predicted to form the classical cloverleaf structure (Supplemental Figs. S15, S16).

Deletion breakpoints are located in mtDNA regions with high folding potentials

We performed a sliding-window analysis of the folding potentials (100-nt windows with 1 nt of overlap) throughout the entire mitochondrial genome (L- and H- strands) to capture the folding energy of all of the possible conformational transitions in which every mtDNA position is involved (Supplemental Fig. S18, S19). The mean ΔG value in the 16,569 100-nt mtDNA windows is -5.504 kcal/mol (standard deviation of 3.387) for the L-strand and -6.366 kcal/mol (standard deviation of 3.403) for the H-strand. There are marked variations in the folding energies across the mtDNA, with sudden increases and decreases in the ΔG values (Supplemental Fig. S19). The mtDNA region with the highest folding potential is the WANCY cluster of tRNAs, matching the second-highest peak in the number of 5' breakpoints (and the 3rd of all breakpoints).

To investigate how the local DNA sequence environment might contribute to the formation of deletions, we extracted and folded all of the 100-nt segments from both the H- and L-strands that enclosed a 5' or 3' breakpoint as the midpoint (Fig. 6A). The code used to identify each region is composed of a number ('5' or '3') according to the type of breakpoint and a letter ('H' or 'L') for the mtDNA strand. The mean observed free energy values were -5.66 (5L), -5.93 (3L), -6.58 (3H) and -7.01 (5H) kcal/mol (Supplemental Fig. S20).

The distribution of free energies of folding along the mtDNA considering all breakpoint areas is represented in Figures 6B, 6C and Supplemental Figures S21, S22. In order to test if there is an association between the number of deletion breakpoints and the folding energy of the breakpoint area, we compared the two mtDNA segments where 5' and 3' breakpoints are more frequent with their upstream and downstream flanking segments with the same length. In both cases, we found that there is a significantly higher number of breakpoints and folding potential (more negative ΔG values) in the target region than in both flanking segments (Fig. 7; Supplemental Fig. S23). For instance, the mean ΔG value (-6.76 kcal/mol) of the sliding windows with midpoints from positions 7,401 to 8,200 (the hotspot of 5' breakpoints upstream and within *COX2*) is significantly lower than the estimated for its upstream (mean ΔG = -6.18 kcal/mol; *p*-value = 4.21×10^{-5}; Student's *t*-test) and downstream (mean ΔG = -4.26 kcal/mol; *p*-value < 1.00×10^{-17}; Student's *t*-test) regions. A significant difference was also found between the 97-nt segment defined by the hairpin element around O_L and the adjacent *tRNA-Cys* gene (positions 5,730 to 5,826) and their flanking regions, for both ΔG and the number of 5' breakpoints parameters (data not shown). Similarly, the 16,001-16,100 mtDNA region (around the 16,071 hotspot) has a significantly higher number of 3' breakpoints (*n* = 201) than its upstream (*n* = 22; *p*-value = 7.05×10^{-3}; Student's *t*-test) and downstream (*n* = 2; *p*-value = 2.78×10^{-3}; Student's *t*-test) regions, together with a significantly higher folding potential (Fig. 7; Supplemental Fig. S23).

We next detailed the relationship between the folding potential of each 100-nt mtDNA sliding window and the number of breakpoints that it harbours. For this purpose, we estimated the number of breakpoints in the window midpoint position, i.e., only position 50 in each 100-nt window was considered. The distribution of 100-nt windows according to the number of breakpoints at the midpoint position shows that a total of 1,114 window midpoint positions have at least one deletion breakpoint (6.7%). In general, windows with more deletion breakpoints have a significantly higher folding potential. For example, windows with more than 8 reported breakpoints have an average folding energy (mean = -11.15 kcal/mol) significantly different (*p*-value = 1.71×10^{-9}; Student's *t*-test) from that of windows with no breakpoints (mean = -5.54 kcal/mol) (Fig. 8). Similarly, windows with 4 to 8 breakpoints in the midpoint positions have an average folding energy (mean = -8.14 kcal/mol) that is significantly different (*p*-value = 6.07×10^{-3}; Student's *t*-test) from that of windows with no breakpoints (mean = -5.54 kcal/mol). The average folding energy of windows with 1, 2 or 3 breakpoints are not significantly different from windows with no breakpoints.

Discussion

Our analyses strongly suggest an important role for non-B DNA structures in mtDNA deletions. Several alterations to the canonical B-form of DNA might occur in the mitochondrial genome with no clear biological function as the simple outcome of particular DNA sequence patterns. However, other structures are important for regulation of replication and transcription (7-11). The structures with no clear biological function vary according to alterations in the primary DNA sequence in a random way and are only removed by purifying selection if their formation interferes with any relevant genomic function. In contrast, those structures with functional relevance are probably maintained under strong selective pressures as previously suggested (8;35).

The non-random genomic distribution of the deletion breakpoints indicates that the root cause of mtDNA vulnerability to breakage resides either in specific characteristics of the local DNA sequence environment or in higher-order features of the genomic architecture. We found that the number of deletion breakpoints in a particular mtDNA position is associated with the folding capacity of the region where it occurs. The genomic regions with more breakpoints have a significant higher folding potential than regions with a low number of breakpoints (Figs. 7, 8). In agreement with this observation, we detected several mtDNA breakage hotspots (e.g., mtDNA positions 16,071, 7,402, 5,787 or 15,435), with cases of different deletions sharing the same breakpoint at one end (Figs. 1, 2; Supplemental Figs. S4, S5). This site-specific hypermutability is a well-known feature of base substitutions within mammalian mtDNA (54). Similarly, our data shows that site-specific breakpoint hotspots exist in the mitochondrial genome, as demonstrated by the significant difference of two orders of magnitude observed between the highest frequencies of breakage in real and simulated data (Supplemental Fig. S7). There is now abundant evidence showing that non-B DNA is more prone to DNA breaks than B-DNA (39-43). For example, the formation of secondary structure intermediates between DNA ends at translocation or gross deletion breakpoints is common in human inherited diseases and cancer (55) and non-B DNA-forming sequences are enriched in breakpoints of copy number variations (56) and chromosomal rearrangements (40-42). It is possible that such extruded bases are more prone to breakage by mechanical or chemical stress. Such conformations can also be the template for the action of trans-acting factors, such as structure-specific nucleases (57).

Since the first reports of mtDNA deletions, scientists have noticed the presence of non-B DNA structures encompassing or in the vicinity of the breakpoints. Just one year after the first description of deletions in human mtDNA (58), Schon et al. (1989) called the attention to the fact that the human mtDNA contains long regions with the potential to form ‘bent DNA’, including around and within the 13-nt repeats that flank the 4,977-bp ‘common

deletion' (8,470–8,482 to 13,447–13,459). The authors noticed that polypyrimidine tracts and AT-rich regions around breakpoints may render such regions susceptible to the formation of single-stranded DNA on supercoiling. Consistent with these observations, we found that the 3' breakpoint of the 'common deletion' is one of the most frequent breaking sites (position 13,447) of mtDNA and occurs within or near a stable hairpin (Fig. 2; Supplemental Fig. S9). No hairpin element has been detected associated with the 5' breakpoint of the 'common deletion', although it has been shown by two-dimensional gel electrophoresis that this genomic region exhibits retarded mobility due to the putative formation of bent DNA structures (31;32).

In the first description of an autosomal disorder causing multiple mtDNA deletions (38), a hotspot for deletion formation was identified near the D-loop 3' termini (at position 16,068 to 16,079). This region is now recognised as a preferential location of mtDNA breakage in both pathological and non-pathological conditions (30;59;60). In their pioneering study, Zeviani and colleagues identified two stable hairpins around this 3' breakpoint hotspot. We have recently discovered that these hairpins are part of a highly conserved 93-nt cloverleaf-like structure (35). In fact, more than one quarter (201 out of 753) of all of the 3' breakpoints occur at this cluster of hairpins, most of them at the 8-nt terminal loop (Figs. 2, 3A). The preference for breakage at single-stranded regions exposed by stable stem elements is also common to other mtDNA regions (Fig. 2; Supplemental Fig. S8, S9). The deletion hotspot around position 16,071 is located near the trinucleotide stop point (16,104-16,106) that is associated with the premature arrest of the H-strand synthesis that forms a three-stranded DNA structure known as D-loop or displacement loop (61) (Fig. 4A). Its location outside of the three-stranded D-loop structure, but still very close to its 3' end, points to a possible relationship between the process of deletion formation and the functional/structural features of the D-loop. Indeed, the formation of the D-loop induces superhelical tension to the mtDNA in solution, at least in *Xenopus laevis* oocytes (62), which might cause distortions of the canonical B-DNA. The complete elucidation of the function(s) of the D-loop under normal physiological conditions will help to understand such potential relationships. Its involvement in mtDNA organisation, segregation and replication have been hypothesised (63;64). The mtDNA breakage hotspot at the D-loop 3' end is not only a feature of human mtDNA. It has been repeatedly found in the homologous region of the mtDNA of transgenic mice with mitochondrial disease (48;65;66). We found that all breaks reported between the *CYTB* gene and the control region of this transgenic mice (48) occur in hairpin elements (Fig. 3). As suggested by Tyynismaa et al. (2007), our data also indicate that the general mechanism underlying multiple deletions with site-specific breakpoints is not due to the primary sequence itself but to its secondary structure or functional location.

The association between hairpin structures and breakpoints is also clear in the WANCY cluster of tRNAs that surrounds the main O_L (Fig. 4B). Our different computational analyses demonstrated that this region stands out as having the highest folding potential of the mitochondrial genome (Figures 6B, 6C and Supplemental Figures S21, S22), with all of the reported 5' deletion breakpoints occurring in hairpin elements (Fig. 4B; Supplemental Fig. S12). The genomic instability at the WANCY cluster is also perceptible when comparing vertebrate mitochondrial genomes: it is a hotspot of gene order rearrangements by tandem duplication and random loss of genes (67). In fact, the breakpoints of such rearrangements observed in different phylogenetic groups are often thought to involve hairpin elements (68).

Our large dataset revealed that deletion breakpoints are overrepresented in tRNA genes (Fig. 5; Supplemental Fig. S14). The structural elements of tRNA molecules, which are well known for playing important biological roles, explain the high folding potential observed in the DNA regions encoding them (e.g., ΔG = -5.1 kcal/mol predicted for the *tRNA-Ser(UCN)* gene). It is therefore likely that the formation of secondary structures at tRNA genes, for instance, during transcription or replication when the mtDNA is single-stranded, may contribute to the formation of the deletions, as previously suggested (33;34). The ability of tRNA genes to mediate genomic rearrangements is well documented in the mitochondrial and other genomes (68-70). A notable example detected in our study is the breakpoint-prone *tRNA-Leu(UUA/G)* gene, which has a significantly higher number of breakpoints than its flanking genes (Fig. 5). Although it is located in the minor arc, it surrounds one of the top ten mtDNA breakage hotspots, at position 3,263 (Fig. 2).

The breaking of the DNA is only one of the many factors that shape the distribution of mtDNA deletions. There are several constraints to the formation of circular deleted mtDNA molecules and their subsequent proliferation in cells that might explain why genomic regions with very high folding potentials (and thus mutational prone) are devoid of detectable breakpoints. It is likely that many breakage sites in mtDNA are not detected just because they are not involved in the formation of circular deleted mtDNA molecules with an efficient replication capacity. Although other constraints might act on mtDNA deletions, the presence of homology at breakpoints and the removal or replication origins are believed to significantly influence their distribution (28-30).

The presence of homology at the edges of mtDNA deletions, including both perfect and imperfect short repeats, is a well-known feature of this type of genomic rearrangement (30;33;37;71). In particular, two 13-nt direct repeats were found associated with most human mtDNA deletions (30). However, an important breakthrough was recently made by Guo et al. (28) by showing that deletion breakpoints coincide with distant segments of mtDNA that are capable of forming stable imperfect duplexes with each other, rather than with the presence of perfect repeats, as previously assumed. In other words, 5' and 3' deletion breakpoints tend

to have large regions of partial or interrupted homology with, at least, 100 nucleotides (28), which is possibly related with the joining of the distant mtDNA regions rather than with the breakage of the DNA strand. Therefore, only those breakage sites near regions with long stretches of high sequence homology are detected in mtDNA deletions. This feature might explain why some putative breakpoint-rich areas are not detected in deletions. For instance, the *tRNA-Pro* gene (devoid of breakpoints) is locate in a region that apparently forms unstable duplexes as suggested by the white or clear 'stripes' around position 16,000 that run across the matrix of free energies made by Guo et al. (28). It is possible that breakpoints in the *tRNA-Pro* gene remain undetectable just because they do not participate in the formation of circular deleted mtDNA molecules. The development of accurate methods to directly detect breakage sites in the mtDNA will clarify this issue.

The mtDNAs without an origin of replication have a limited replication capacity that influences their likelihood of propagation and detection. This feature clearly influences the distribution of mtDNA deletions, which are mostly located inside the major arc (88.8% of the cases) without removing any origin of replication (Fig. 1). In fact, we only identified 71 deletions (9.43% of all cases) without the O_L and one deletion without the O_H (0.13% of all cases), as defined by the strand-asynchronous model of mtDNA replication (14;72), even taking into account that we have designed our database to include deletions that were detected in post-mitotic tissues at low levels (e.g., from Kajander et al. (59)), the detection of which does not depend on an efficient replication capacity. It is probably because of its location inside the minor arc near O_H that rRNA and some tRNA genes do not display a high number of breakpoints, meaning that large deletions with a breakpoint in this area are likely to remove an origin of replication. Moreover, the profile of folding energy across the rRNA genes is not different from that observed in the rest of the genome (Supplemental Figure S19). The highest folding potential is reached at the end of the *RNR1* gene, with a ΔG of -18.19 kcal/mol (L-strand sequence) below the folding capacity of the WANCY region (ΔG of -20.56 kcal/mol; window with midpoint at position 5,732). The rare presence of breakpoints in the rRNA genes might also be related with the mTERF-mediated molecular process for rRNA synthesis that postulated the looping-out of the rDNA region (73). The high rate of rRNA synthesis and the formation of a large DNA loop between termination and initiation sites might somehow interfere with the deletion process. Further work is necessary to uncover all the constraints to the distribution of deletions. Nevertheless, we were able to find a clear association with non-B DNA conformations even taking into account that many breakage hotspots might remain undetectable.

Various hypotheses have been advanced in recent years to explain the formation of mtDNA deletions, each one giving a different prominence to the roles of DNA replication (60;71), recombination (33;37) and repair (29) in the deletion process. Despite the intense

debate, there is no compelling evidence unequivocally demonstrating which mechanism (if only one) generates deletions in the various physiological situations. The data presented here indicate that, whatever the process underlying deletions, non-B DNA conformations (intra-strand hairpins and cloverleaf-like elements) should be considered an important piece in the complex puzzle of mitochondrial genomic rearrangements. Unfortunately, our poor knowledge of the *in vivo* dynamics and organisation of mtDNA makes it very difficult to understand exactly how such alternative DNA structures cause deletions.

It was recently suggested that mtDNA deletions are most likely to occur during repair of damaged mtDNA (29). The formation of non-B DNA might trigger this process by increase the rate of single and double-strand lesions (39-43). In alternative, several models posit that mtDNA deletions are a replication-associated phenomenon (60;71;74). Previous *in vivo* and *in vitro* experiments have shown that non-B DNA acts as a preferential pausing site for DNA polymerases (75;76), which may be an obstacle to fork progression or a target for nucleolytic attack, thus permitting DNA breakage and deletion formation (77-80). The preferential location of 5' and 3' breakpoints in the large arc between O_H and O_L suggests that the replicative process might somehow influence the generation of deletions. Consistent with this assumption, mice expressing a defective mtDNA polymerase display elevated replication pausing and breakage at fragile sites near O_L (65;81). A different mechanism suggests that the formation of a transient DNA triple helix in pyrimidine-rich sequences might guide the slipped mispairing of the replication complex, causing mtDNA rearrangements (36). In addition, the accumulation of multiple mtDNA deletions in individuals with defects in the replicative helicase Twinkle and DNA polymerase γ have been used to support the idea that deletion formation might be induced by replication stalling (60). This process might be related to the breakpoint hotspot in the cluster of hairpins near the D-loop 3' end (Fig. 2A), a putative replication fork barrier.

The formation of hypermutable non-B DNA in certain regions of the mitochondrial genome might indeed be the link between the high incidence of deletions in individuals with defects in proteins that move, organise or replicate the DNA (which collectively control the mtDNA topology). The inefficient activity of the altered versions of such DNA-interacting proteins might induce considerable changes in the topology and supercoiled state of mtDNA, whose tensional stress might be released by the formation of non-B DNA and the breakage of DNA. In fact, even the normal movement of RNA and DNA polymerases through the mtDNA may generate regions of superhelical tension and other topological alterations, which may be associated with the genesis of mtDNA rearrangements (82). More detailed biochemical and computational studies are needed to verify all of these conjectures. Although the mechanisms remains elusive, our analyses suggest that DNA structure-induced genomic instability seems to be at the heart of the mtDNA deletion process.

Supplementary Data

Supplementary Data are available at NAR online: Supplementary figures S1-S23.

Funding

This work was supported by the Portuguese Foundation for Science and Technology (FCT) grant SFRH/BPD/44637/2008 and research project PTDC/CVT/100881/2008 to FP. IPATIMUP is an Associate Laboratory of the Portuguese Ministry of Science, Technology and Higher Education and is partially supported by FCT.

Figures

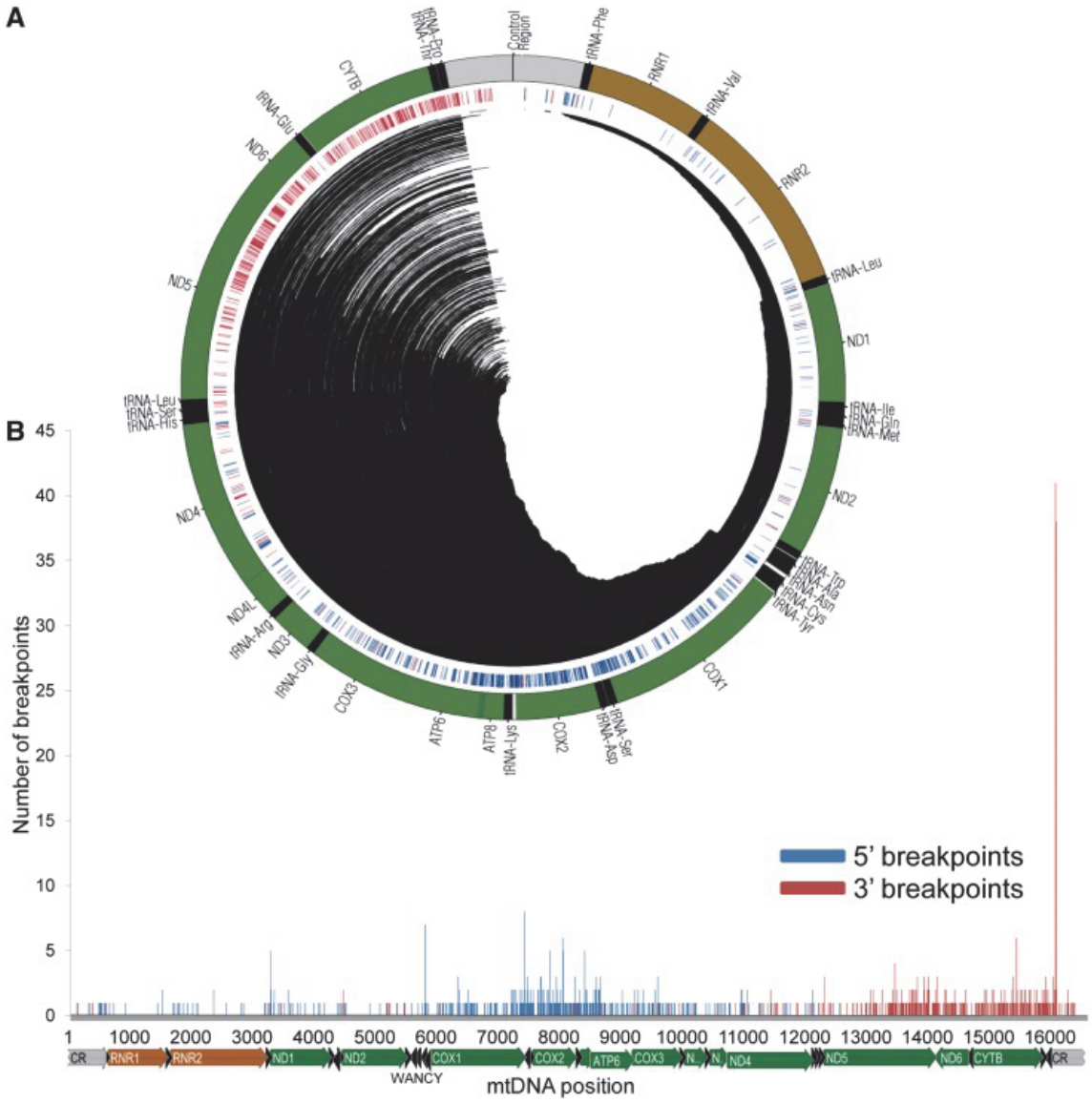


Figure 1: Deletion breakpoints are not randomly distributed throughout the mitochondrial genome. (A) A circular representation of the human mtDNA with annotated tRNA (black), rRNA (brown) and protein-coding (green) genes (outer track). The central track depicts the location of 5′- (blue) and 3′- (red) breakpoints. The central lines indicate the deleted region in the 753 reported cases. (B) The distribution of 5′- (blue bars) and 3′- (red bars) deletion breakpoints in the human mtDNA. The locations of the mitochondrial genes are shown below the x-axis.

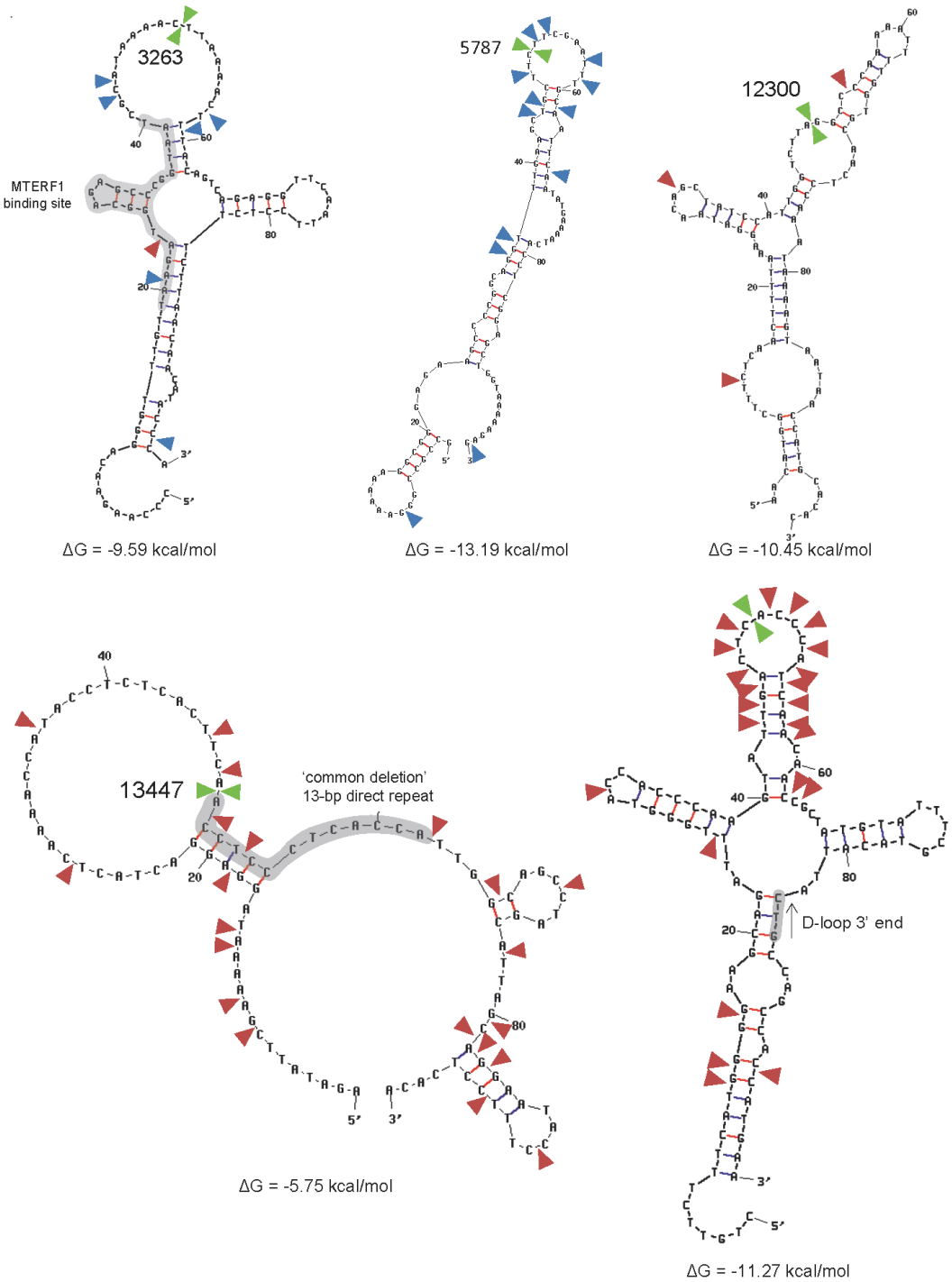


Figure 2: The most frequent deletion breakpoints occur within or near predicted hairpins. Five of the most frequent breakpoint sites (mtDNA positions 3263, 5787, 12300, 13447 and 16071) are indicated by a green arrow in the predicted structure (L-strand) of the 100-nt flanking region (breakpoints were used as window midpoints). The blue and red arrows indicate less frequent 5′- and 3′-breakpoints, respectively. Highlighted in grey are the binding sites of the MTERF1, the 13-nt direct repeat at the 3′-breakpoint of the 'common deletion' and the D-loop 3′-terminus.

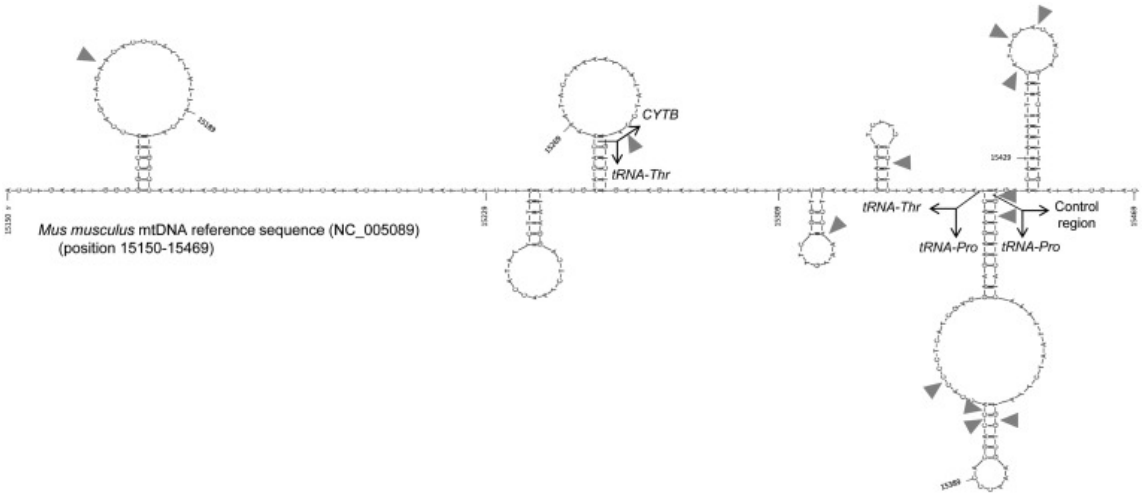


Figure 3: Breakage sites in the mouse mitochondrial genome are associated with hairpin elements. All of the deletion breakpoints described in transgenic mice with mitochondrial disease (yellow arrows) are indicated in the *Mus musculus* mtDNA L-strand reference sequence (NC_005089), from position 15,150 to 15,469. The secondary structures were obtained in the mfold-util software v4.6.

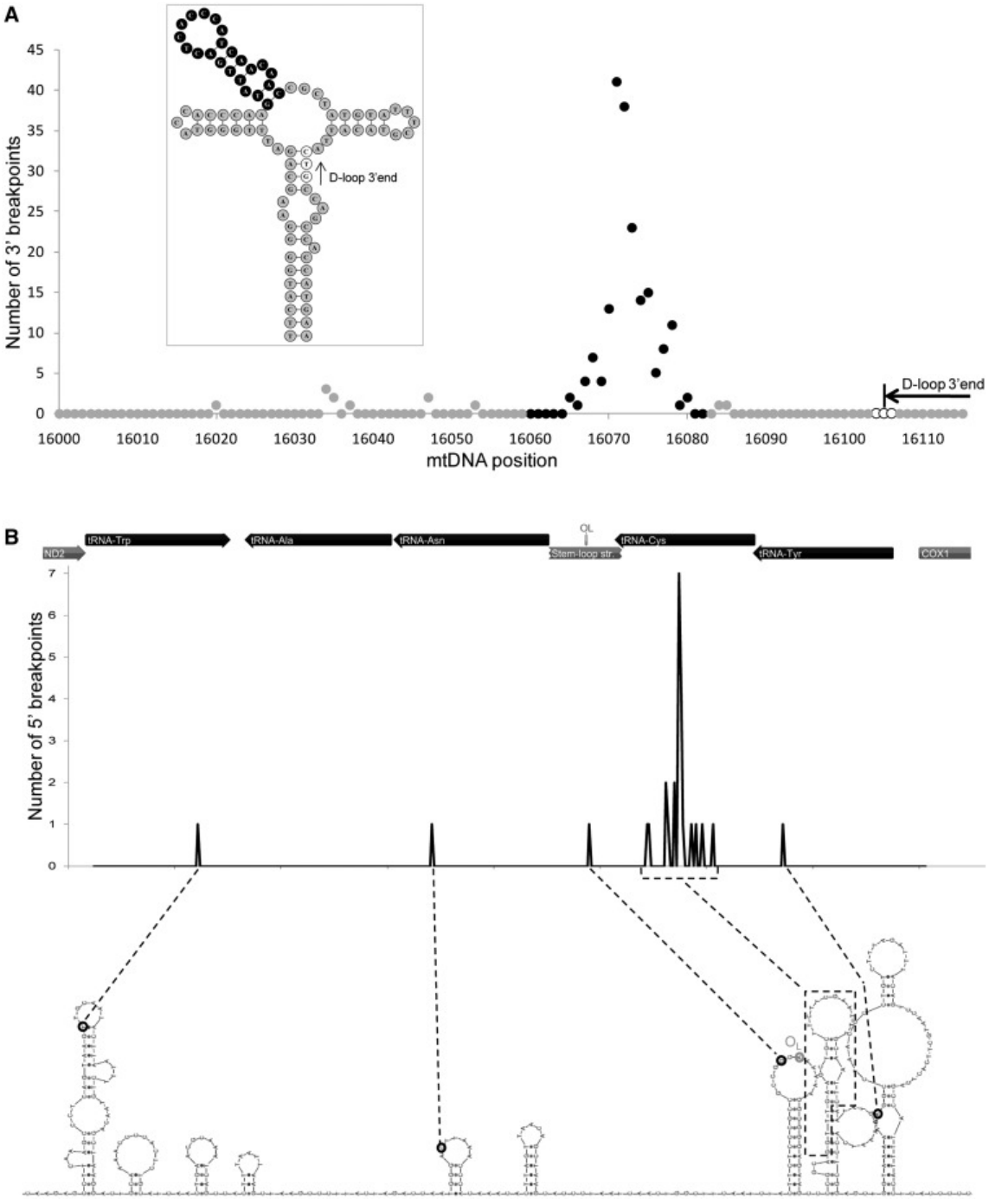


Figure 4: Two stable clusters of hairpins are hotspots for mtDNA breakage. (A) A total of 189 reported 3' breakpoints occur in the central hairpin (black circles) of a large cloverleaf-like structure (enclosed image) predicted for a 93-nt stretch (positions 16,028 to 16,120) of the control region near the Proline tRNA (L-strand). Inside the hairpin, 144 breakpoints (19% of all of the 3' breakpoints) are located on the 8-nt terminal loop. This deletion hotspot is located near the trinucleotide stop point (16,104-16,106; white circles) for the premature arrest of the H-strand synthesis responsible for forming a three-stranded DNA structure known as the displacement loop (D-loop). (B) All of the 5' deletion breakpoints (*n* = 27) identified in the WANCY cluster of tRNAs are located in hairpin elements (L-strand). Most of them (*n* = 23) are located in a single stem-loop element predicted for the *tRNA-Cys* gene, downstream of the stem-loop structure that is associated with the origin of L-strand replication (O_L).

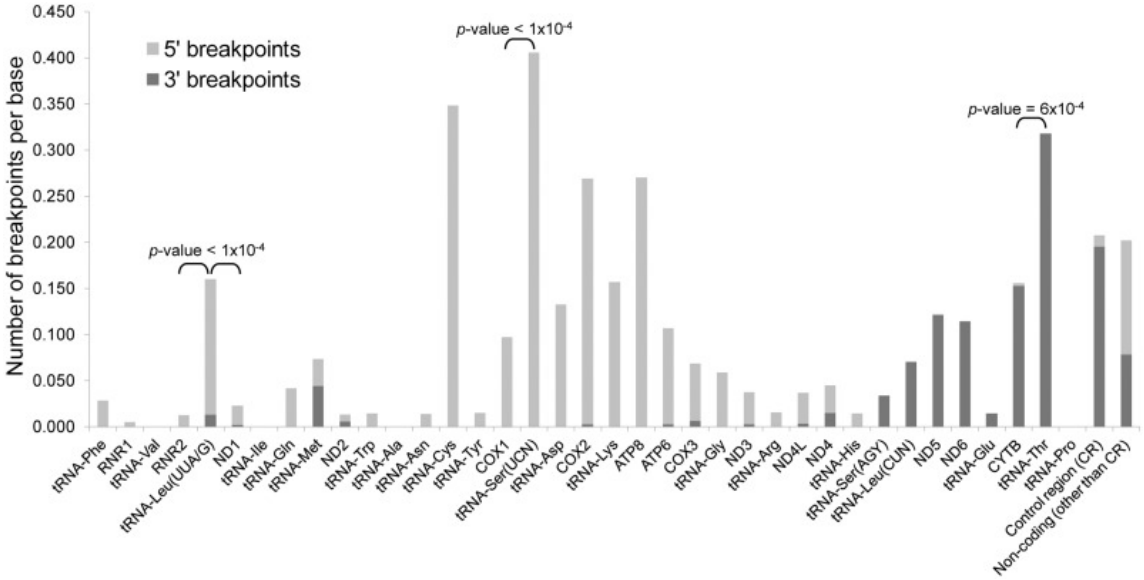


Figure 5: Mitochondrial tRNA genes are hotspots for mtDNA breakage. The graph displays the number of deletion breakpoints per base (number of breakpoints/region length) according to the coding features of the mitochondrial genome (5' and 3' breakpoints in light and dark grey, respectively). The significance of the difference between the number of breakpoints in some tRNA and their flanking genes is shown (two-sided *p*-values; Student's *t*-test).

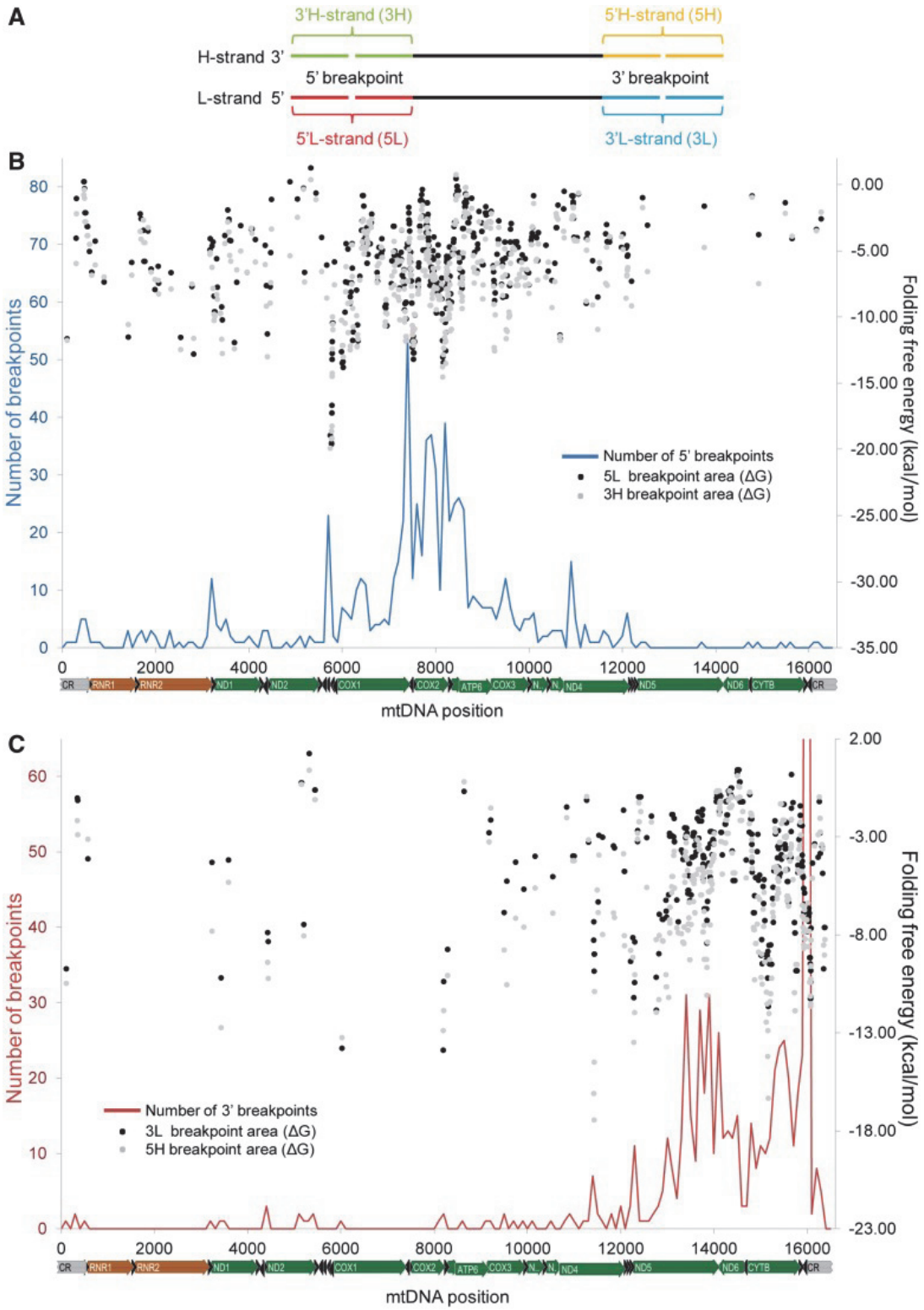


Figure 6: Genome-wide distribution of folding potentials in the breakpoint areas of mtDNA deletions. (A) The nomenclature and schematic representation of the location of the genomic region enclosing 5' and 3' deletion breakpoints. (B, C) Black and grey dots indicate the free energy of folding (kcal/mol) of 100-nt windows around the 5' (B) and 3' (C) breakpoints (black and grey dots for L-strand and H-strand segments, respectively). The blue (B) and red (C) lines indicate the distribution of 5' and 3' breakpoints, respectively (measured in 100-nt sliding windows with an overlap of 1 nt). The peak at the 16,071 hotspot reaches 201 deletions but is not completely shown to facilitate the visualisation of smaller peaks.

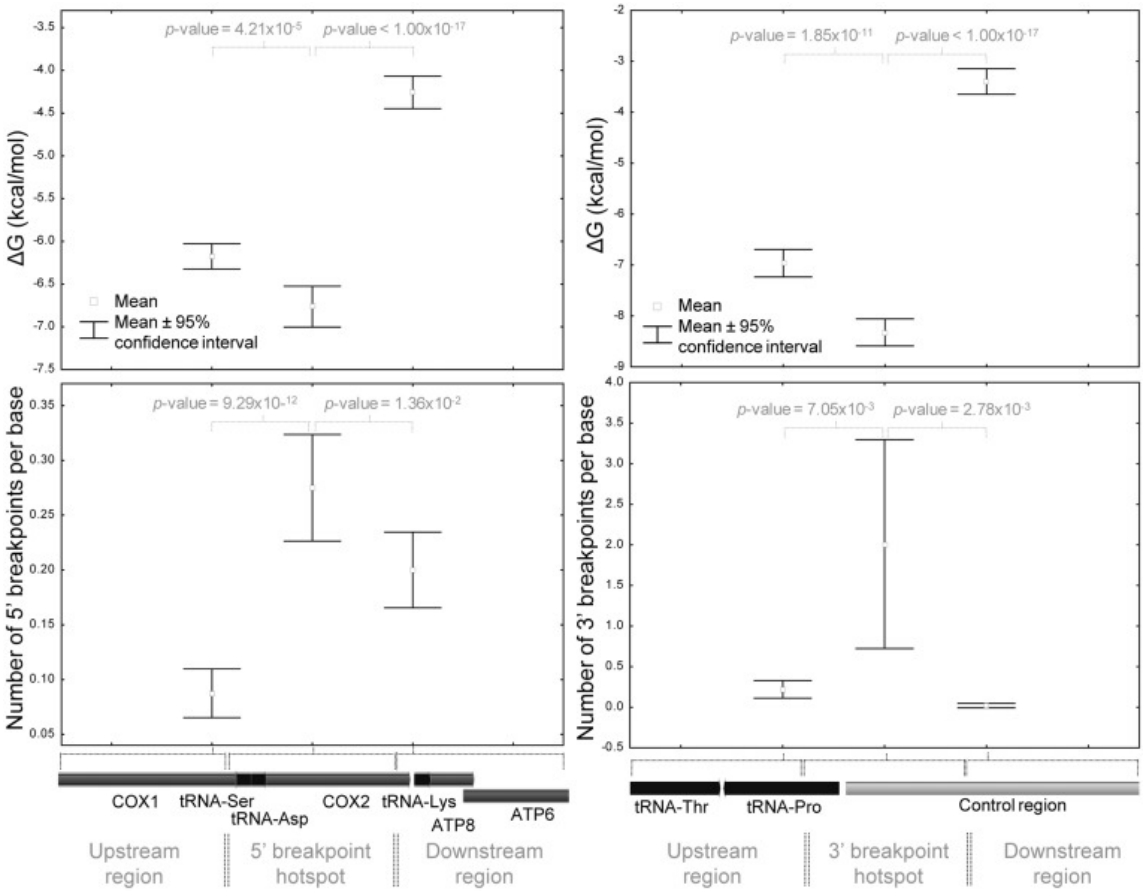


Figure 7: The main hotspots of mtDNA breakage have higher folding potentials than adjacent regions. The two mtDNA segments where 5′- and 3′-breakpoints are more frequent (positions 7401–8200 and 16 001–16 100, respectively) were compared with their upstream and downstream flanking segments. We estimated for each segment the mean number of breakpoints per base (and the 95% confidence interval for the mean) and the average folding potential of the 100-nt windows with a midpoint position in that region. There is a significant higher folding potential (more negative Δ*G* values; top graphs) and higher number of breakpoints per base (bottom graphs) in the hotspot regions than in their flanking segments. The results of the statistical tests (Student's *t*-test; two-sided *P*-values) to evaluate the differences in means (Δ*G* values and number of breakpoints) between adjacent regions are indicated.

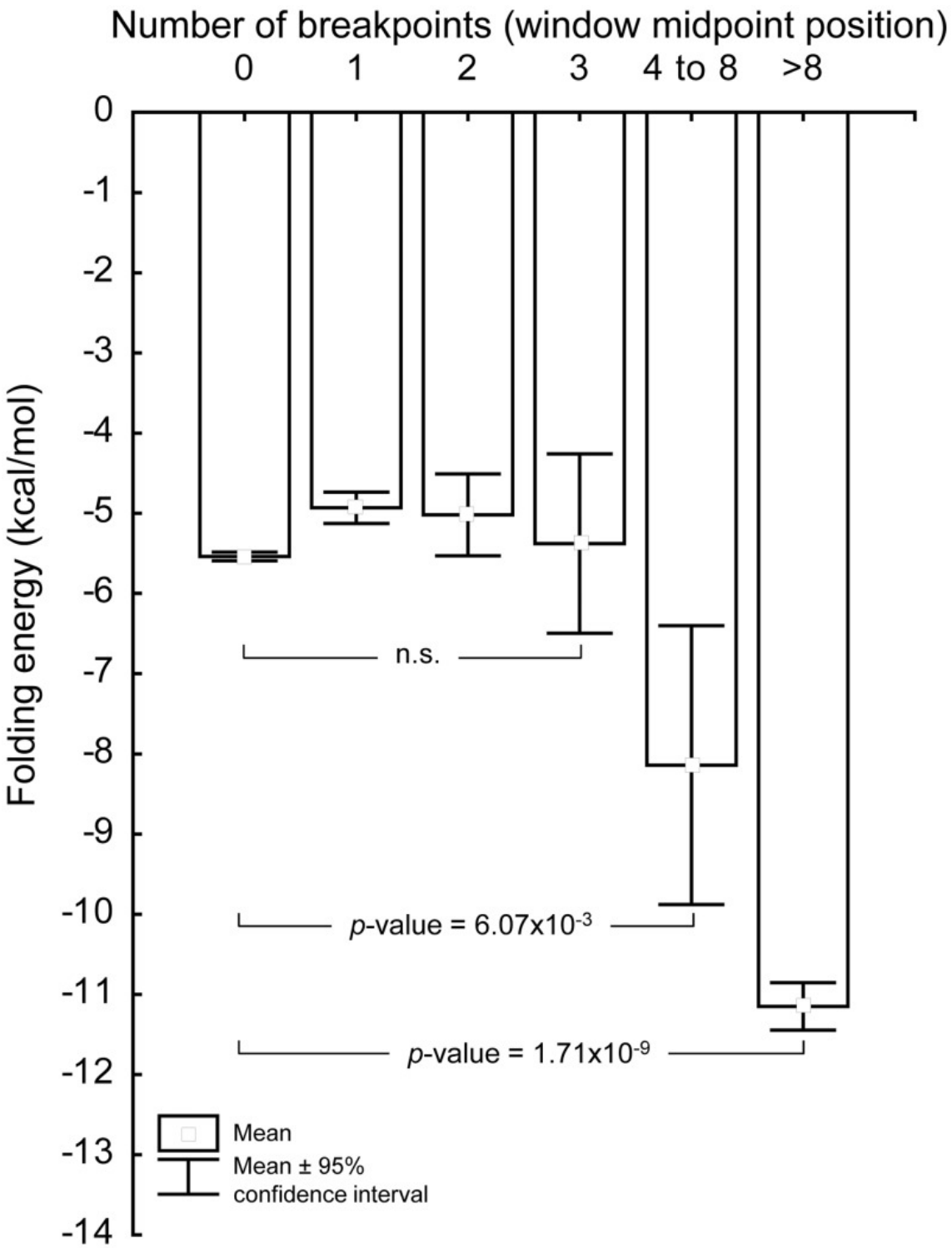


Figure 8: Deletion breakpoints are located in mtDNA regions with high folding potentials. The bar chart depicts the relationship between the mean folding potential (ΔG values) and the number of breakpoints in the midpoint position of 100-nt sliding windows covering the entire mitochondrial genome. The results of the statistical tests (Student's *t*-test; two-sided *p*-values) shows that windows with more breakpoints have a higher folding potential than windows where breakpoints are rare.

References

1. Cooper,D.N., Bacolla,A., Ferec,C., Vasquez,K.M., Kehrer-Sawatzki,H. and Chen,J.M. (2011) On the sequence-directed nature of human gene mutation: the role of genomic architecture and the local DNA sequence environment in mediating gene mutations underlying human inherited disease. Hum Mutat., 32, 1075-1099.
2. Wells,R.D. (2007) Non-B DNA conformations, mutagenesis and disease. Trends Biochem. Sci, 32, 271-278.
3. Zhao,J., Bacolla,A., Wang,G. and Vasquez,K.M. (2010) Non-B DNA structure-induced genetic instability and evolution. Cell Mol Life Sci, 67, 43-62.
4. Bates,A. and Maxwell,A. (2005) DNA topology. Oxford University Press, Oxford.
5. Mirkin,S.M. (2001) DNA Topology: Fundamentals. Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd, Chichester.
6. Wells,R.D., Dere,R., Hebert,M.L., Napierala,M. and Son,L.S. (2005) Advances in mechanisms of genetic instability related to hereditary neurological diseases. Nucleic Acids Res, 33, 3785-3798.
7. Fuste,J.M., Wanrooij,S., Jemt,E., Granycome,C.E., Cluett,T.J., Shi,Y., Atanassova,N., Holt,I.J., Gustafsson,C.M. and Falkenberg,M. (2010) Mitochondrial RNA polymerase is needed for activation of the origin of light-strand DNA replication. Mol Cell, 37, 67-78.
8. Hixson,J.E., Wong,T.W. and Clayton,D.A. (1986) Both the conserved stem-loop and divergent 5'-flanking sequences are required for initiation at the human mitochondrial origin of light-strand DNA replication. J Biol Chem, 261, 2384-2390.
9. Ngo,H.B., Kaiser,J.T. and Chan,D.C. (2011) The mitochondrial transcription and packaging factor Tfam imposes a U-turn on mitochondrial DNA. Nat Struct. Mol Biol, 18, 1290-1296.
10. Rubio-Cosials,A., Sidow,J.F., Jimenez-Menendez,N., Fernandez-Millan,P., Montoya,J., Jacobs,H.T., Coll,M., Bernado,P. and Solà,M. (2011) Human mitochondrial transcription factor A induces a U-turn structure in the light strand promoter. Nat Struct. Mol Biol, 18, 1281-1289.
11. Yakubovskaya,E., Mejia,E., Byrnes,J., Hambardjieva,E. and Garcia-Diaz,M. (2010) Helix unwinding and base flipping enable human MTERF1 to terminate mitochondrial transcription. Cell, 141, 982-993.
12. Lilley,D.M. (1980) The inverted repeat as a recognizable structural feature in supercoiled DNA molecules. Proc Natl. Acad. Sci U. S A, 77, 6468-6472.
13. Panayotatos,N. and Wells,R.D. (1981) Cruciform structures in supercoiled DNA. Nature, 289, 466-470.
14. Clayton,D.A. (2003) Mitochondrial DNA replication: what we know. IUBMB Life, 55, 213-217.
15. Falkenberg,M., Larsson,N.G. and Gustafsson,C.M. (2007) DNA replication and transcription in mammalian mitochondria. Annu Rev Biochem., 76, 679-699.
16. Larsson,N.G. (2010) Somatic mitochondrial DNA mutations in mammalian aging. Annu Rev Biochem., 79, 683-706.
17. Dayn,A., Malkhosyan,S. and Mirkin,S.M. (1992) Transcriptionally driven cruciform formation *in vivo*. Nucleic Acids Res, 20, 5991-5997.
18. Postow,L., Crisona,N.J., Peter,B.J., Hardy,C.D. and Cozzarelli,N.R. (2001) Topological challenges to DNA replication: conformations at the fork. Proc Natl. Acad. Sci U. S A, 98, 8219-8226.
19. Fisher,R.P., Lisowsky,T., Parisi,M.A. and Clayton,D.A. (1992) DNA wrapping and bending by a mitochondrial high mobility group-like transcriptional activator protein. J Biol Chem, 267, 3358-3367.
20. Brown,T.A., Tkachuk,A.N. and Clayton,D.A. (2008) Native R-loops persist throughout the mouse mitochondrial DNA genome. J Biol Chem, 283, 36743-36751.
21. Burger,G., Gray,M.W. and Lang,B.F. (2003) Mitochondrial genomes: anything goes. Trends Genet, 19, 709-716.

22. Bender,A., Krishnan,K.J., Morris,C.M., Taylor,G.A., Reeve,A.K., Perry,R.H., Jaros,E., Hersheson,J.S., Betts,J., Klopstock,T. et al. (2006) High levels of mitochondrial DNA deletions in substantia nigra neurons in aging and Parkinson disease. Nature Genetics, 38, 515-517.
23. Kraytsberg,Y., Kudryavtseva,E., McKee,A.C., Geula,C., Kowall,N.W. and Khrapko,K. (2006) Mitochondrial DNA deletions are abundant and cause functional impairment in aged human substantia nigra neurons. Nature Genetics, 38, 518-520.
24. Cree,L.M., Samuels,D.C. and Chinnery,P.F. (2009) The inheritance of pathogenic mitochondrial DNA mutations. Biochimica et Biophysica Acta-Molecular Basis of Disease, 1792, 1097-1102.
25. DiMauro,S. and Hirano,M. (2003) Mitochondrial DNA Deletion Syndromes. In Pagon,R., Bird,T., Dolan,C. and Stephens,K. (eds.), GeneReviews. University of Washington, Seattle.
26. Stewart,J.B., Freyer,C., Elson,J.L. and Larsson,N.G. (2008) Purifying selection of mtDNA and its implications for understanding evolution and mitochondrial disease. Nature Reviews Genetics, 9, 657-662.
27. Taylor,R.W. and Turnbull,D.M. (2005) Mitochondrial DNA mutations in human disease. Nature Reviews Genetics, 6, 389-402.
28. Guo,X., Popadin,K.Y., Markuzon,N., Orlov,Y.L., Kraytsberg,Y., Krishnan,K.J., Zsurka,G., Turnbull,D.M., Kunz,W.S. and Khrapko,K. (2010) Repeats, longevity and the sources of mtDNA deletions: evidence from 'deletional spectra'. Trends Genet, 26, 340-343.
29. Krishnan,K.J., Reeve,A.K., Samuels,D.C., Chinnery,P.F., Blackwood,J.K., Taylor,R.W., Wanrooij,S., Spelbrink,J.N., Lightowlers,R.N. and Turnbull,D.M. (2008) What causes mitochondrial DNA deletions in human cells? Nature Genetics, 40, 275-279.
30. Samuels,D.C., Schon,E.A. and Chinnery,P.F. (2004) Two direct repeats cause most human mtDNA deletions. Trends in Genetics, 20, 393-398.
31. Hou,J.H. and Wei,Y.H. (1996) The unusual structures of the hot-regions flanking large-scale deletions in human mitochondrial DNA. Biochem. J, 318 (Pt 3), 1065-1070.
32. Hou,J.H. and Wei,Y.H. (1998) AT-rich sequences flanking the 5'-end breakpoint of the 4977-bp deletion of human mitochondrial DNA are located between two bent-inducing DNA sequences that assume distorted structure in organello. Mutat. Res, 403, 75-84.
33. Mita,S., Rizzuto,R., Moraes,C.T., Shanske,S., Arnaudo,E., Fabrizi,G.M., Koga,Y., DiMauro,S. and Schon,E.A. (1990) Recombination via flanking direct repeats is a major cause of large-scale deletions of human mitochondrial DNA. Nucleic Acids Res, 18, 561-567.
34. Nishigaki,Y., Marti,R. and Hirano,M. (2004) ND5 is a hot-spot for multiple atypical mitochondrial DNA deletions in mitochondrial neurogastrointestinal encephalomyopathy. Hum Mol Genet, 13, 91-101.
35. Pereira,F., Soares,P., Carneiro,J., Pereira,L., Richards,M.B., Samuels,D.C. and Amorim,A. (2008) Evidence for variable selective pressures at a large secondary structure of the human mitochondrial DNA control region. Molecular Biology and Evolution, 25, 2759-2770.
36. Rocher,C., Letellier,T., Copeland,W.C. and Lestienne,P. (2002) Base composition at mtDNA boundaries suggests a DNA triple helix model for human mitochondrial DNA large-scale rearrangements. Mol Genet Metab, 76, 123-132.
37. Schon,E.A., Rizzuto,R., Moraes,C.T., Nakase,H., Zeviani,M. and DiMauro,S. (1989) A direct repeat is a hotspot for large-scale deletion of human mitochondrial DNA. Science, 244, 346-349.
38. Zeviani,M., Servidei,S., Gellera,C., Bertini,E., DiMauro,S. and DiDonato,S. (1989) An autosomal dominant disorder with multiple deletions of mitochondrial DNA starting at the D-loop region. Nature, 339, 309-311.
39. Bacolla,A., Jaworski,A., Larson,J.E., Jakupciak,J.P., Chuzhanova,N., Abeysinghe,S.S., O'Connell,C.D., Cooper,D.N. and Wells,R.D. (2004) Breakpoints of gross deletions coincide with non-B DNA conformations. Proc. Natl. Acad. Sci U. S. A, 101, 14162-14167.

40. Darmon,E., Eykelenboom,J.K., Lincker,F., Jones,L.H., White,M., Okely,E., Blackwood,J.K. and Leach,D.R. (2010) E. coli SbcCD and RecA control chromosomal rearrangement induced by an interrupted palindrome. Mol Cell, 39, 59-70.
41. Inagaki,H., Ohye,T., Kogo,H., Kato,T., Bolor,H., Taniguchi,M., Shaikh,T.H., Emanuel,B.S. and Kurahashi,H. (2009) Chromosomal instability mediated by non-B DNA: cruciform conformation and not DNA sequence is responsible for recurrent translocation in humans. Genome Res, 19, 191-198.
42. Raghavan,S.C., Swanson,P.C., Wu,X., Hsieh,C.L. and Lieber,M.R. (2004) A non-B-DNA structure at the Bcl-2 major breakpoint region is cleaved by the RAG complex. Nature, 428, 88-93.
43. Wang,G., Christensen,L.A. and Vasquez,K.M. (2006) Z-DNA-forming sequences generate large-scale deletions in mammalian cells. Proc Natl. Acad. Sci U. S A, 103, 2677-2682.
44. Mani,R.S. and Chinnaiyan,A.M. (2010) Triggers for genomic rearrangements: insights into genomic, cellular and environmental influences. Nat Rev Genet, 11, 819-829.
45. Fan,L. and Yao,Y.G. (2011) MitoTool: a web server for the analysis and retrieval of human mitochondrial DNA sequence variations. Mitochondrion, 11, 351-356.
46. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res, 32, 1792-1797.
47. Knight,R., Maxwell,P., Birmingham,A., Carnes,J., Caporaso,J.G., Easton,B.C., Eaton,M., Hamady,M., Lindsay,H., Liu,Z. et al. (2007) PyCogent: a toolkit for making sense from sequence. Genome Biol, 8, R171.
48. Tyynismaa,H., Mjosund,K.P., Wanrooij,S., Lappalainen,I., Ylikallio,E., Jalanko,A., Spelbrink,J.N., Paetau,A. and Suomalainen,A. (2005) Mutant mitochondrial helicase Twinkle causes multiple mtDNA deletions and a late-onset mitochondrial disease in mice. Proc Natl. Acad. Sci U. S A, 102, 17687-17692.
49. Markham,N.R. and Zuker,M. (2008) UNAFold: software for nucleic acid folding and hybridization. Methods Mol Biol, 453, 3-31.
50. SantaLucia,J., Jr. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc Natl. Acad. Sci U. S A, 95, 1460-1465.
51. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. Journal of Molecular Biology, 288, 911-940.
52. Krzywinski,M., Schein,J., Birol,I., Connors,J., Gascoyne,R., Horsman,D., Jones,S.J. and Marra,M.A. (2009) Circos: an information aesthetic for comparative genomics. Genome Res, 19, 1639-1645.
53. Sadikovic,B., Wang,J., El-Hattab,A., Landsverk,M., Douglas,G., Brundage,E.K., Craigen,W.J., Schmitt,E.S. and Wong,L.J. (2010) Sequence homology at the breakpoint and clinical phenotype of mitochondrial DNA deletion syndromes. PLoS One, 5, e15687.
54. Galtier,N., Enard,D., Radondy,Y., Bazin,E. and Belkhir,K. (2006) Mutation hot spots in mammalian mitochondrial DNA. Genome Res, 16, 215-222.
55. Chuzhanova,N., Abeysinghe,S.S., Krawczak,M. and Cooper,D.N. (2003) Translocation and gross deletion breakpoints in human inherited disease and cancer II: Potential involvement of repetitive sequence elementsin secondary structure formation between DNA ends. Human Mutation, 22, 245-251.
56. Conrad,D.F., Pinto,D., Redon,R., Feuk,L., Gokcumen,O., Zhang,Y.J., Aerts,J., Andrews,T.D., Barnes,C., Campbell,P. et al. (2010) Origins and functional impact of copy number variation in the human genome. Nature, 464, 704-712.
57. Bzymek,M., Saveson,C.J., Feschenko,V.V. and Lovett,S.T. (1999) Slipped misalignment mechanisms of deletion formation: *In vivo* susceptibility to nucleases. Journal of Bacteriology, 181, 477-482.
58. Holt,I.J., Harding,A.E. and Morgan-Hughes,J.A. (1988) Deletions of muscle mitochondrial DNA in patients with mitochondrial myopathies. Nature, 331, 717-719.

59. Kajander,O.A., Rovio,A.T., Majamaa,K., Poulton,J., Spelbrink,J.N., Holt,I.J., Karhunen,P.J. and Jacobs,H.T. (2000) Human mtDNA sublimons resemble rearranged mitochondrial genomes found in pathological states. Human Molecular Genetics, 9, 2821-2835.

60. Wanrooij,S., Luoma,P., van,G.G., van,B.C., Suomalainen,A. and Spelbrink,J.N. (2004) Twinkle and POLG defects enhance age-dependent accumulation of mutations in the control region of mtDNA. Nucleic Acids Res, 32, 3053-3064.

61. Kasamatsu,H., Robberson,D.L. and Vinograd,J. (1971) A novel closed-circular mitochondrial DNA with properties of a replicating intermediate. Proc Natl. Acad. Sci U. S A, 68, 2252-2257.

62. Callen,J.C., Tourte,M., Dennebouy,N. and Mounolou,J.C. (1983) Changes in D-loop frequency and superhelicity among the mitochondrial DNA molecules in relation to organelle biogenesis in oocytes of Xenopus laevis. Exp. Cell Res, 143, 115-125.

63. Antes,A., Tappin,I., Chung,S., Lim,R., Lu,B., Parrott,A.M., Hill,H.Z., Suzuki,C.K. and Lee,C.G. (2010) Differential regulation of full-length genome and a single-stranded 7S DNA along the cell cycle in human mitochondria. Nucleic Acids Res, 38, 6466-6476.

64. Holt,I.J., He,J., Mao,C.C., Boyd-Kirkup,J.D., Martinsson,P., Sembongi,H., Reyes,A. and Spelbrink,J.N. (2007) Mammalian mitochondrial nucleoids: organizing an independently minded genome. Mitochondrion, 7, 311-321.

65. Bailey,L.J., Cluett,T.J., Reyes,A., Prolla,T.A., Poulton,J., Leeuwenburgh,C. and Holt,I.J. (2009) Mice expressing an error-prone DNA polymerase in mitochondria display elevated replication pausing and chromosomal breakage at fragile sites of mitochondrial DNA. Nucleic Acids Res, 37, 2327-2335.

66. Srivastava,S. and Moraes,C.T. (2005) Double-strand breaks of mouse muscle mtDNA promote large deletions similar to multiple mtDNA deletions in humans. Hum Mol Genet, 14, 893-902.

67. San,M.D., Gower,D.J., Zardoya,R. and Wilkinson,M. (2006) A hotspot of gene order rearrangement by tandem duplication and random loss in the vertebrate mitochondrial genome. Mol Biol Evol, 23, 227-234.

68. Stanton,D.J., Daehler,L.L., Moritz,C.C. and Brown,W.M. (1994) Sequences with the potential to form stem-and-loop structures are associated with coding-region duplications in animal mitochondrial DNA. Genetics, 137, 233-241.

69. Cantatore,P., Gadaleta,M.N., Roberti,M., Saccone,C. and Wilson,A.C. (1987) Duplication and remoulding of tRNA genes during the evolutionary rearrangement of mitochondrial genomes. Nature, 329, 853-855.

70. Juhling,F., Putz,J., Bernt,M., Donath,A., Middendorf,M., Florentz,C. and Stadler,P.F. (2011) Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. Nucleic Acids Research.

71. Shoffner,J.M., Lott,M.T., Voljavec,A.S., Soueidan,S.A., Costigan,D.A. and Wallace,D.C. (1989) Spontaneous Kearns-Sayre/chronic external ophthalmoplegia plus syndrome associated with a mitochondrial DNA deletion: a slip-replication model and metabolic therapy. Proc Natl. Acad. Sci U. S A, 86, 7952-7956.

72. Robberson,D.L., Kasamatsu,H. and Vinograd,J. (1972) Replication of mitochondrial DNA. Circular replicative intermediates in mouse L cells. Proc Natl. Acad. Sci U. S A, 69, 737-741.

73. Martin,M., Cho,J.Y., Cesare,A.J., Griffith,J.D. and Attardi,G. (2005) Termination factor-mediated DNA loop between termination and initiation sites drives mitochondrial rRNA synthesis. Cell, 123, 1227-1240.

74. Buroker,N.E., Brown,J.R., Gilbert,T.A., Ohara,P.J., Beckenbach,A.T., Thomas,W.K. and Smith,M.J. (1990) Length heteroplasmy of sturgeon mitochondrial DNA - an illegitimate elongation model. Genetics, 124, 157-163.

75. Viguera,E., Canceill,D. and Ehrlich,S.D. (2001) Replication slippage involves DNA polymerase pausing and dissociation. Embo Journal, 20, 2587-2595.

76. Weaver,D.T. and Depamphilis,M.L. (1982) Specific Sequences in Native Dna That Arrest Synthesis by Dna Polymerase-Alpha. Journal of Biological Chemistry, 257, 2075-2086.

77. Bierne,H., Ehrlich,S.D. and Michel,B. (1997) Deletions at stalled replication forks occur by two different pathways. Embo Journal, 16, 3332-3340.

78. Bzymek,M. and Lovett,S.T. (2001) Evidence for two mechanisms of palindrome-stimulated deletion in Escherichia coli: Single-strand annealing and replication slipped mispairing. Genetics, 158, 527-540.

79. Glickman,B.W. and Ripley,L.S. (1984) Structural intermediates of deletion mutagenesis - a role for palindromic DNA. Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences, 81, 512-516.

80. Madsen,C.S., Ghivizzani,S.C. and Hauswirth,W.W. (1993) In-vivo and in-vitro evidence for slipped mispairing in mammalian mitochondria. Proceedings of the National Academy of Sciences of the United States of America, 90, 10409.

81. Trifunovic,A., Wredenberg,A., Falkenberg,M., Spelbrink,J.N., Rovio,A.T., Bruder,C.E., Bohlooly,Y., Gidlof,S., Oldfors,A., Wibom,R. et al. (2004) Premature ageing in mice expressing defective mitochondrial DNA polymerase. Nature, 429, 417-423.

82. Tengan,C.H. and Moraes,C.T. (1998) Duplication and triplication with staggered breakpoints in human mitochondrial DNA. Biochim. Biophys. Acta, 1406, 73-80.

10. Publication IV: Molecular Dynamics Simulations on Tetranucleotide Short Tandem Repeats Small Hairpins

Molecular Dynamics Simulations on Tetranucleotide Short Tandem Repeats Small Hairpins

João Carneiro[1,2*], Filipe Pereira[1], António Amorim[1,2], Raquel Silva[1], Luísa Azevedo[1,2], Irina Moreira[2,3], Maria João Ramos[2,3]

[1] Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), R. Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

[2] Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

[3] REQUIMTE/Departamento de Química e Bioquímica, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal

* To whom correspondence should be addressed:
João Carneiro
joaomiguelsov@gmail.com
Irina Moreira
irina.moreira@fc.up.pt

Abstract

Microsatellites, or short tandem repeats (STRs) represent about 3% of the human genome, with most of them located in non-coding regions. Although it was assumed that they might not have a biological function, recent studies showed the influence of STRs on gene expression, recombination, and maintenance of chromatin spatial organization. It remains to be determined how the number of repeats vary according to the stepwise mutation model (SMM), and what is the role that strand-slippage replication mechanisms have on the STR mutational process. In order to address these questions, we have performed molecular dynamics simulations in tetranucleotide motif STRs from the Y chromosome (Y-STRs). Our results indicate that different forms of hairpins are predicted without a clear association to any specific initial constraints, which suggests that formation of small hairpins can occur in any region of the STR and depends on specific conditions of the STR region (base composition, counterions, and water distribution). The hairpin presence (stem base pairing and loop) was associated to electronegative pockets of Na+ that were present near hairpins stems or loops (<10 Å). Each STR showed a specific folding potential which will influence STRs length variation resulting from replication errors (e.g., strand-slippage replication mechanism).

Introduction

Non-coding regions represent almost 99% of the human genome[1]. It is now believed that they are important to the organization and evolution of the genome [2,3]. The ENCODE project [4-10] has recently demonstrated the importance of non-coding regions as regulatory elements. For instance, histones modifications and chromatin accessibility in these regions have important consequences to replication process and transcription [4,5,10]. A particular class of non-coding DNA corresponds to microsatellites or short tandem repeats (STRs), which represent nearly 3% of the human genome [1] and are assumed to be neutrally evolving. This class of elements is particularly dynamic and highly polymorphic due to a relatively high mutation rate that induces variation in the number of units of the repeated motif [11,12]. Therefore, STRs are widely used as genetic markers in population genetics, in forensics, and evolutionary studies [13-17]. In addition, they are also associated to several human diseases [18,19] that are believed to result from replication errors, repair errors and genomic instability, such as fragile X syndrome, Huntington's disease, myotonic dystrophy, and various types of spinocerebellar ataxias [20-25] The factors that influence the STR mutation rate have been under extensive investigation, including the variation in repeat number and size, base composition, flanking sequences, recombination, sex and age of the individual [16,19,26-32]. The variation in STRs length can produce alterations in the state of the chromatin conformation or promote the insertion of transposable elements, which will have consequences to genome architecture[33]. In this respect, the SMM has been proposed to explain the length heterogeneity in STRs [19,34]. The SMM model posits that most mutations correspond to single repeat unit additions or removals [35,36] and is based on the premise that the size of alleles is virtually unlimited and there is independence between mutation rate and repeat number.

Molecular dynamics (MD) approaches to describe DNA helical systems, single-stranded systems and other particular conformations have been used to better understand DNA behaviour in different environments [37-41]. MD has some limitations related with the force-field used to represent the molecular systems (e.g., quantum effects in some biological processes like changes in chemical bonding, cannot be modelled), the estimate of interatomic potentials, the size of the systems and the time scales that are appropriate to mimic biologically important processes [42-46].

The Unified Nucleic Acid Folding and hybridization package (UNAFold) [47] is a well-established software that uses experimentally determined thermodynamic parameters to predict non-canonical RNA and DNA secondary structures. There is a clear association of these secondary structure predictions and biochemical processes [48-50]. We performed a MD simulation to mimic the DNA conformations in Y-STR taking into consideration the

UNAFold secondary structure predictions and their implications in replication mechanisms [48,51,52].Our new approach designed to explain the heterogeneity observed in STRs at the structural level demonstrates that fold/unfold states can occur frequently enough to increase the potential formation of hairpin structures in single-stranded STRs. The folding states observed in every conformational state where associated with specific environment of water and counterions [41,53-55]. Overall, our simulations demonstrate that hairpin conformations can occur and be maintained during 12 ns dynamics, are associated to specific water and counterion distributions, and can adopt supercoiled conformations. The results also suggest that the motif nucleotide composition has great influence in the maintenance of these small hairpins, and different molecular systems can maintain the initially constrained stems and loops.

Methods

Y-STR Homology Search in NDB

We searched the Nucleic Acid Database (NDB) to retrieve Protein Data Bank (PDB) format files homologous to each Y-STR available at the Y-STR National Institute of Standards and Technology (NIST) database [56] and from Ballantyne et al. [11] (Supplementary Material Table SI1). A tailor-made python script was developed for this purpose. We started the search using the minimal allele length (1 repeat), and then increased the search until we reached the maximum allele length (e.g., DYS19A search was done with allele lengths ranging from 2 to 19 repeats). Identity scores for each search were saved.

UNAFold Secondary Structure Predictions and NAB 3D Structures

The accuracy of Unified Nucleic Acid Folding and hybridization package (UNAFold) was tested by using a database of 22 nuclear magnetic resonance (NMR) spectroscopy experimentally-determined single-stranded DNA structures (Table 1) with small hairpins. We compared these 3D structures with the secondary structure prediction of small hairpins of UNAFold.

The UNAFold software was then used to predict the secondary structure for each Y-STR. We have used NABpy python class (NABpy Python Supplementary File) to predict UNAFold secondary structures of all Y-STRs in NIST database using the following parameters: temperature of 37 Celsius (°C), sodium ion concentration (mol/L) of 0.05 and magnesium ion concentration of 0.002. Temperature and ion conditions were selected to mimic a PCR system at the starting temperature [57,58]. We used the PCR parameters as a

simplified model that represents the most important variables of the DNA single-stranded molecules present in *in vivo* replication process.

DNA strands of each STR were built representing 3D molecular systems of single-stranded and constrained conformations. The small hairpin presence and/or nucleotide base pairing were observed during MD. The initial constrained hairpin structures were predicted using thermodynamics parameters calculated by UNAFold and described elsewhere [47,51]. The base pairing geometry of experimentally determined structures implemented in NAB was used to build the stems of the 3D single-stranded DNA presenting small hairpins. In order to build the 3D models we selected three of the most common tetranucleotide Y-STRs from the complete NIST database considering: a) UNAFold secondary structure prediction with small hairpins (Figure 1A, Table 2 and Supplementary Material Table SI1); b) different nucleotide composition; c) motif length of 4 nucleotides; d) different middle range allele size (DYS19A: 15 repeats, DYS391: 10 repeats, DYS531: 11 repeats). We produced for each STR several DNA macromolecules (5' – 3' strands) using the nucleic acid builder [NAB [59]] as implemented in NABpy: one structure with single-stranded Y-STR (SS), one with Y-STR UNAFold predicted small hairpin secondary structure constraints (UF) and others considering UNAFold predicted small hairpin with identical loop and same base pair connections in stem, but not located in the region of STR predicted by UNAFold (UFR) (Figure 1B and Supplementary Material Table SI2). The single-stranded DNA molecule was used as control for each STR. The UFR molecules were built to determine if the presence of small hairpins is dependent of the STR region where they occur. We also analysed each STRs (DYS19A, DYS391 and DYS531) considering 11 repeats molecular systems (single-stranded and small hairpins UNAFold constrained DNA), to ascertain if STRs with same length maintain the small hairpins during MD. We neutralized the molecular DNA systems with Na+ and solvated them with explicit TIP3P water molecules that extended 10 Å from any edge of the box to the DNA atoms. This specific solvent environment was selected to simulate PCR conditions.

Explicit solvent MD of the single-stranded and UNAFold predicted DNA conformations resembling an *in-vitro* system were performed for 12 ns. The computational analysis to test the stem and/or loop presence in Y-STR was performed taking in consideration the secondary structure (small hairpins) formation implications in replication mechanisms [48,51,52].

Molecular Dynamics Simulation

To model the UNAFold prediction of stable secondary structures, we performed MD simulations of the UNAFold prediction and linear molecules of DNA. For each model we created the topology and simulation parameters files to run the MD during 12 ns using the parm99SB Amber force field [60]. The systems were initially energy minimized to remove bad

contacts by steepest descent followed by conjugate gradient algorithms. Subsequently, they were subjected to 4 ns of heating procedure (in NVT ensemble) in which the temperature was gradually raised to 300 K, followed by 8 ns runs in the NPT ensemble. The Langevin thermostat [61,62] was used and the electrostatic interactions were calculated by using the particle mesh Ewald (PME) method [63]. Bond lengths involving hydrogen atoms were constrained to their equilibrium values using the SHAKE algorithm [64]. The equations of motion were integrated with a two fs time-step and the non-bonded interactions were truncated with a 10 Å cutoff. All computations for the simulation of the generated Y-STR molecular models using a solvated box TIP3P with 10 ångström (Å) of side were completed and analysed as described in Figure SI2 in Supplementary Material. Each nucleotide is referred using AmberTools [59,65] nomenclature: Adenine (DA), 3' Adenine (DA3), Thymine (DT), 5' Thymine (DT5), Guanine (DG) and Cytosine (DC).

Root-Mean-Square-Deviation (RMSD) Graphs and End to End Distances

We analysed the MD stability by calculating the RMSD values for all atoms and backbone atoms. All graphs were generated parsing the RMSD file with NABpy and plotting the data with Matplotlib (http://matplotlib.org/).The last 4 ns of every simulation were condensed into one trajectory file: Trajectory file 1 (RMS1) with ≈350 snapshots; the waters and ions were not removed. We also determined the distance of 5' C1' carbon atoms (first residue) and 3' C1' carbon atoms (last residue) and end to end distances graphical outputs of each molecular dynamics simulation in RMS1.

Hairpin States and H-Bond Trajectory Analysis

The hairpin states in each molecular model were analysed by considering three regions of low free energy: 1) the native form (with 6–8 hydrogen bonds and around four stacked bases); 2) a partially folded state characterized by two hydrogen bonds and two stacked bases, and 3) the fully unfolded form with no native hydrogen bonds and a variable number of stacked bases [41]. The process of small hairpins formation is elsewhere described [41,66-68] and can result from different folding mechanisms. Hairpin formation involving loop nucleation of one base pairing connection followed by second base pairing connection results from fluctuations between closed and open hairpin states [68].

All functions to perform hydrogen bonds (H-bonds) analysis of trajectory files using AmberTools were implemented in NABpy. We detected the base-pairing occurring in 3D Y-STR small hairpins by using ptraj in AmberTools. We used a mask for ptraj hbond function considering all possible hydrogen bond acceptors in each nucleotide (DA@N1, N3, N7; DT@O2, O4; DG@O6, N3; DC@N1, N3 , O2) and all possible hydrogen bond donors (DA@N6 :DA@H61, DA@N6 :DA@H62; DT@N3 :DT@H3; DG@N2 :DG@H21; DG@N2 :DG@H22,DG@N1 :DG@H1; DC@N4 :DC@H41,DC@N4 :DC@H42) with a cut-off of 5 Å. Non-canonical H-bonds between base pairs were considered in this analysis. The water (cut-off of 3 Å) and ions (cut-off of 5 Å) H-bonds were also analysed. We considered as hydrogen bond acceptors the phosphate groups (@O1P; @O2P) and oxygen atoms (@O3'; @O4'; @O5').). All results were compiled using NABpy and exported to Microsoft Excel. Graphical outputs were generated in STATISTICA v10 [69].

X3DNA Analysis and Curves+ Analysis

We calculated an average coordinate file (saved as PDB) for each Y-STR model molecular dynamics for the RMS1. The X3DNA [70] software was used to determine the canonical base pairs connections occurring in the average PDB. The Curves+ software [71] was used to calculate the nucleotide parameters (backbone, inter and axis base pair parameters) of each Y-STR MD simulation of RMS1. Coordinate files (PDBs) corresponding to each snapshot were created and then analysed with Curves+. The results were compiled for each nucleotide in each STR using NABpy. The results were exported to Microsoft Excel and graphical outputs were generated in STATISTICA v10.

Results

Sequence Homology Search

The sequence homology search for all NIST database Y-STRs performed over Nucleic Acid Database [56] did not give any alignment (identity values>0.5) with NDB sequences, considering only one tetranucleotide repeat. For each single repeat unit we obtained alignment sequence identity values of 0.393 for DYS19A (TAGA repeat), 0.286 for DYS391 (TCTA repeat), and 0.386 for DYS531 (AAAT repeat). The alignment sequence identity for more than one repeat, using the reported allele range at NIST, was near zero for all the searched lengths, which means that no similar PDB data was available in NDB to build our models using homology. We could not obtain reliable PDBs (at least sequence identity >0.5) to build the three dimensional (3D) macromolecular models for the selected Y-STR sequences. Therefore, we built all molecular simulation models (Figure 1A and 1B, Supplementary Material Table SI1 and Figure SI1) using UNAFold (stem base pairing constrains), NABpy and NAB from the AMBER package.

UNAFold Secondary Structure Predictions

Since UNAFold predictions present some limitations to accurately determine some types of structures we have performed an accuracy test for small hairpin predictions. Considering the NMR 3D structures analysed to test the UNAFold accuracy we obtained an overall result that gives a high performance for the small hairpin prediction of UNAFold software. We have used a database of experimentally determined NMR structures of 22 single-stranded DNA with hairpins (Table 1 and Figure 2, http://ndbserver.rutgers.edu/index.html) to test the accuracy of UNAFold to predict hairpin structures. We obtained 20 accurate loop predictions (90.9%), 18 accurate stem and loop prediction (81.8%). The only structures that UNAFold was unable to predict were 1EL2 and 1ELN because these are complex 3D structures that fold over themselves.

We obtained UNAFold small hairpin structures predictions for all 36 Y-STRs described in Supplementary Material Table SI1. These small hairpin structures were predicted to be just one per STR and located at 5' end regions (Figure 1A), even when considering different allele lengths. The detected fold/unfold states of all tested models were predicted to be the main cause of allele heterogeneity (Figure 1C) since we had avoid the effect of recombination, which is one of the factors involved in the generation of variation, by using STRs located in the non-recombining Y chromosome.

Molecular Dynamics: Limitations Considering the Initial Starting Structures

The MD simulations analysed here present limitations considering the starting structures used. Nevertheless current force fields can be used to model unusual DNA structures [60,72]. The fact that small hairpins with 10 nucleotides and two base pairing connections are been analysed, reduces the probability of errors related with changes in chemical bonding. Other limitation was the 12 ns time scale used in the MD simulations. The time scale for hairpin folding via MD (1000 ns) [41] observed in other single-stranded DNA systems was not achieved in this analysis but we properly modelled the single-stranded small hairpins in each STR for 12 ns. We could observe the presence, absence and maintenance of small hairpins in MDs and determine if there were high fluctuations in conformations for each molecular model tested. The methodology used here can be used for large time scales that can mimic biologically important processes.

Molecular dynamics: RMSD

The RMSD graphs of each Y-STR presented a similar behaviour for the backbone atoms and for all the atoms (Graphs SI1-SI29 of Supplementary Material and Figure 3). This result indicates that, for the analysed Y-STRs, changes in conformation during 12 ns are consistent both for DNA backbone structure and peripheral atoms. In all cases, a major conformational switch occurred in the first 4 ns with RMSD values from 12 to 18 Å. Stabilization of the RMSD values is achieved consistently in the last 4 ns of trajectory results (Figure 3). We also observed slight variations in conformation stabilization when comparing the UNAFold prediction conformations occupying other regions of the same STR. The RMSD values variation for different STRs is not as high as expected when comparing the single-stranded models. We expected to have more differences considering the various repeats in each STR (DYS19A: TAGA, DYS391: TCTA; DYS531: AAAT) and the different allele lengths tested (DYS19A: 15 repeats, DYS391: 10 repeats; DYS531: 11 repeats). Even considering this fact the RMSD values were very high for all the molecular models tested. Different models for the same STR resulted in different RMSD variations, and therefore in very different conformations. This result is in agreement with the fact that the molecular systems were subjected to different constraints in the beginning of each simulation.

Molecular Dynamics: End to end distances

To test the 5' and 3' conformational changes of each Y-STR molecular model during the MD simulation, we analysed the distance between the C1 atoms of the first and last nucleotide of the STR. The results are summarized in Table SI3 of Supplementary Material. The average distance between the Y-STR strand ends for the DYS19A models is 45.14 Å (maximum average of 56.49 Å; minimum average of 34.85 Å). The first nucleotide (1DT) to last nucleotide (60DA) distance average variation of DYS19A UNAFold predicted structure is 25.26 Å, the lowest when compared with all the other tested models. Nevertheless, the 5' and 3' median distance is sufficiently high (≈45 Å) in all models to consider that the beginning and end of the DYS19A strand is sufficiently distant enough to validate the models (Figure 4 and Supplementary Material Table SI3). The distance variation along last 4 ns of trajectories corroborates these results (Supplementary Material Graphs SI30-SI58).

The DYS391 (40 nucleotide length) average distance between first residue (1-DT5:C1') and last residue (40-DA3:C1') is 41.15 Å. The single-stranded molecular model was one of the systems where the average distance is lower (35.35 Å). The difference of the average distance value comparing the Y-STR DYS19A (60 nucleotides length) with the Y-STR DYS391 (40 nucleotide length) is ≈4 Å. This result is mainly derived from the higher strand length of DYS19A. The average distance (41.36 Å) of the C1' atoms of DYS531 (44 nucleotide length) was in accordance with the DYS391 value that has almost the same length.

The MD simulations are consistent among the three Y-STRs molecular models when the end to end distance between the first residue C1' atoms and the last residue C1' atoms are considered. An approximation of the 5' and 3' ends is not observed during the MD simulations.

Molecular Dynamics: H-Bonds

We found several differences in DYS19A when comparing the alternative Y-STR constraints hypothesis (Supplementary Material Tables SI4-SI14, Graphs SI59-SI69, and Graphs SI88-SI95). We obtained the lowest number of valid H-bond contacts between bases for the single-stranded 3D model of DYS19A Y-STR. During the last 4 ns of the STR MD simulation the number of preserved H-bond connections (canonical or non-canonical nucleotide interaction) was always higher for the molecular models submitted to different forced constraints (203-288 H-bond contacts). These conformational states are very stable at the 5' end of the STR as determined by the UNAFold software. Nevertheless, upon the analyses of simulations where the hairpin was located at 3' end of the STR, the number of H-bond contacts was still higher than the simulation considering a single-stranded Y-STR molecule (Supplementary Material Tables SI4-SI14). In all these cases we obtained localized

H-bonds between nucleotides that were not contiguous (more than two bases distance). Spontaneous formation of several H-bonds (stacking or base-pairing interactions) can occur in DYS19A along the STR, which are mainly isolated base connections. When considering the UNAFold prediction (Table 3), in particular the specific region where constraints were forced in the beginning of simulation, we noticed an elongation of the stem formed between 4<->13 and 5<->12 [only one 4<->12 H-bond in the form N7<->H62<->N6 (acceptor-Hydrogen-donor)]. A new stem has arisen between 5<->12 and 6<->11 (DT<->DA and DA<->DG), which suggests that stem-loops in different positions at different times of the molecular dynamics can arise (Table 3). In the other molecular models tested for DYS19A (Table 4), the H-bond pattern was maintained with localized H-bonds in the regions where the initial constraints were present (stem base pairing and stacking). The two base pairing connections were not maintained in all UFR models tested, and in some cases only one connected base pair was maintained showing a loop nucleation. These results are in agreement with the hairpin closed and open states mechanism that where described by Goddard et al. [68] by comparing different stem-loops conformations using a thermal equilibrium analysis between closed and open conformations.

The highest number of H-bond base pairing in DYS391 was observed in the single-stranded DNA molecular system, while the lowest (almost half of single-stranded DYS391) was in the UNAFold constrained molecule (Figure 5, Supplementary Material Tables SI15-SI22, Graphs SI70-SI77, and Graphs SI96-SI103). This result clearly contrasts with the STR previously analysed. The observed pattern may be explained by the inherent instability of these small hairpins that can fold and unfold in a few ns [41,66,68]. While the DYS19A UNAFold structure maintains localized H-bonds base pairing after initial constraints, DYS391 has the same behaviour described by Orozco and co-workers as a trapped structure in a stable compact but non-native conformation that does not reach native minima [41]. For the DYS391 initially constrained structure (UNAFold prediction) we obtained partial folded form (1 hydrogen bond between nucleobases 4<->12 (DA<->DA); 3 hydrogen bonds between nucleobases 3<->11 (DT<->DT), which is localized one nucleotide away from initial H-bonds constraints). The DYS391 single-stranded molecule presents 6 H-bonds between 4<->10 (DA<->DC) nucleobases with no consecutive nucleobases connections to form a stem. The UFR model with 8<->17 (DT<->DA) and 9<->16 (DT<->DA) base pairing, presented only 2 H-bonds during the MD simulation (occupancy of 18.56 and 6.89). Therefore, the last 4 ns of simulation presented an unfolded conformation (no native H-bonds and variable number of stacked bases). We obtained no H-bonds between initial forced parameters in 16<->25 and 17<->24 DYS391 UFR model. All the other UFR systems presented a stem but the connected nucleobases were not well defined, although the number of H-bonds was always

equal or higher than 4 (4-13 H-bonds) between the initial constrained base pairs or near nucleobases.

We observed the different conformations described for the two previous STRs (unfolded and partially folded) when analysing the DYS531 (11 repeats) models (Supplementary Material Tables SI23-SI32, Graphs SI78-SI87 and Graphs SI104-SI111). The initial H-bonds present in UNAFold predicted conformation were all lost [0 H-bonds between 3<->12 (DA<->DT) and 4<->11 (DT<->DA) nucleobases], suggesting that the motif base composition (AAAT), has great influence in the maintenance and formation of these small hairpins. High number of H-bonds in single-stranded molecular model was observed in the region predicted to have a loop by UNAFold, but did not occurred in DYS391 and DYS19A. In addition, native non-canonical H-bond base pairing between nucleobases 4<->12 (DT<->DT) in the end of simulation was observed. In the UFR systems tested we obtained two (DYS531AAAT11-35b44b-36b43b-UFR and DYS531AAAT11-11b20b-12b19b-UFR) with no initial constrained H-bonds maintained during the molecular dynamics. All the other UFR systems presented H-bond connections between initial base pairing nucleotides and/or near nucleotides, which demonstrates that different constrained systems can maintain the initial constrained hairpins.

As discussed in other articles the H-bond between base pairs can represent significant variations in energetic values (-2 to -3 kcal/mol/H-bond), and if we consider the total interactions of a base pair, the binding energy value ranges from -5 to -47 kcal/mol [73-75]. In DYS19A, we observed different number of H-bonds between the systems, considering the presence or absence of hairpins presence. Some models presented a super-coiled conformation associated with a compact 3D structure (Figure 6) related with relaxation/tension status in different regions of STR (folded regions alternating with stretched regions). The DYS391 free energies (h-bonds contribution presents also significant differences between the models) also presented high changes in free energetic values between models. The DYS391 UNAFold predicted model can be associated to a supercoiled conformation as mentioned in Figure 7. DYS531 presents a similar pattern. The pattern observed here is consistent with the presence of supercoiled DNA of the DYS531 models (Figure 8), except for the single-stranded model.

Molecular Dynamics: Counterions and Water Distribution

We neutralized the DNA systems using Manning's concept for the distribution of counterions in molecular DNA aqueous systems [76]. The anionic phosphates are neutralized by the counterion atmosphere of DNA, which promotes the electrostatic stability to the system. Water activity and ion distribution (composition and concentration) modulate the DNA structure. The water and ion distribution is even more critical considering that conformation abrupt changes (global or local) can occur during molecular dynamics of different types of DNA.

We analysed Na+ distribution in the DYS19A MD simulations and observed that only the UNAFold predicted structure did not present Na+ atoms near DNA strand (5 Å cut-off) with high occupancy values (>30% occupancy). All the other models tested showed occupancy values >30% for one or more residues. The distribution in UNAFold predicted structure molecular dynamics was very uniform (Supplementary Material Graphs SI112-SI122). We also observed an almost perfect fitting straight line by the distance weighted least squares method. When comparing the variation of Na+ distribution through specific regions of this STR in other models, we noticed the existence of high occupancy peaks of residues near the middle of the STR. Thus, the Na+ ions with high occupancy were not evenly distributed along the STR sequence in contrast with the low occupancy ones. In the single-stranded DYS19A the Na+ counterions near residues 32 (DA) and 23 (DG) presented the highest occupancy values with 79.94% and 76.65% respectively. The Na+ counterions that were very close to nucleotides 34 (DA) and 14 (DA) in UNAFold predicted MD, presented lower occupancy values (25.75% and 20.66%, respectively). In general, the single-stranded model presented a total number of possible H-bond between Na+ and DNA strand of 55 (4 with high occupancy) with a mean distance of 4.34 Å, while for UNAFold prediction system the total was 76 possible H-bonds with an equal mean distance but with low occupancy values for all counterions near residues. There was no specific association between initial constrained residues (hairpin stem base pairing) of each model and the residues found 5 Å distance from Na+ counterions in the final of each MD. This indicates that some of the folded regions observed and maintained in the last 4 ns of trajectories are not correlated with any particular Na+ environment considering a 5 cut-off, although we cannot state that Na+ molecules (10 Å cut-off, as described on DNA helical systems by B. Jayaram et al. [76]) can influence the process of fold/unfold of these hairpins. We observed electronegative regions near hairpins of almost all tested models, with three or more Na+ counterions with distances higher than 5 Å from the nucleotides (Figure 6 and 7).The presence of the observed specific counterion environment around stem or loops may be critical to folded/unfolded conformations of STRs.

The behaviour observed in DYS19A was not found in the DYS391 STR tested systems. The occupancy values were globally lower and the position of the closest Na+ atoms to the DNA strand was almost always in the beginning or end region (5' and 3') (Supplementary Material Graphs SI123-SI130). Single-stranded model presented 20 different hydrogen bond interactions between Na+ and nearby nucleotides, while UNAFold prediction model had 31 interactions in the last 4 ns of the MD simulation. The distribution for the high occupancy Na+ atoms of all the molecular systems was mainly unimodal with few high occupancy counterions.

Analysing the results of Na+ distribution for DYS531 (Supplementary Material Graphs SI131-SI140), counterions with low occupancy values were located close to DNA strand for few picoseconds (ps), as previously described for the other STRs. However, we observed an increase of the number of Na+ atoms with high occupancy values (>50%). The single-stranded molecular model presented six high occupancy counterions through the STR strand but not in the 5' end and 3' end regions. The UNAFold prediction system presented only one high occupancy counterion and a high decrease in the total number of Na+ and DNA interactions (18 H-bond interactions in total against 35 of the single-stranded DYS531 STR). These results are related with a medium sized hairpin (between position 15 and 30) observed in the central DYS531 UNAFold predicted model during the molecular dynamics simulation. Although the base pairing H-bonds, already described, were higher for the UNAFold prediction system (177 against a total of 155 H-bonds in the single-stranded system), the Na+ interactions were very low and outside the central region. This behaviour is due to the decreased accessibility of the medium size hairpin of the central region of the DNA strand. In spite of this pattern for Na+ atoms within 5 Å distance from DNA strand, there were electronegative pockets (three or more Na+ atoms) around the fully or partially folded stem-loop structures (Figure 8). Therefore, Na+ distribution around DNA is important in the stabilization of these regions. Possible specific mobile counterions appear near DNA strand during short periods of time resulting in electronegative pockets [37] that can influence the formation of secondary structures (e.g., hairpins). We notice the presence of Na+ counterions, for a few ps, close to initial hydrogen bonding interacting residues that were maintained during the MD (Supplementary Material Graphs SI112-140). The Na+ presence within 5 Å was not detected in regions with stem-loops conformation, but near enough (between 5 Å and 10 Å) to influence folding and unfolding processes. These local conformations may have great relevance for biological mechanism such as replication, where polymerase enzymes [20,21,77,78] do not have total accessibility to interact with the DNA strand because of the folding status of specific STR regions. This can be extended to another already described processes related with replication and disease where formation of

conformational structures plays a role in large duplications of repetitive DNA segments [79,80].

Since the sugar-phospate backbone is especially sensitive to local environment, which means that variations in water distribution can determine significant changes in conformations [81,82], we have also analysed the water distribution around nucleotides. Water interacts with anionic, hydrophilic and hydrophobic constituents of DNA [76]. The distribution of waters molecules around the DYS19A single-stranded DNA (Supplementary Material Graphs SI141-SI151) was uniform in the tested molecular systems and the first hydration shell were maintained around the 3 Å. Water interactions are biased to oxygen-phosphate and oxygen groups, excluding O2 and O6 (Supplementary Material Graphs SI170-SI213). The first water shell layer, taking in consideration the hydrogen bonding, was between 2.8 and 3.0 Å. Some DYS19A molecular systems (e.g., SS, DYS19A-20b29b-21b28b-UFR, DYS19A-24b33b-25b32b-UFR) present water interactions with high occupancy values.

The DYS391 molecular systems also presented a stable first water shell within the 3 Å distance with a global number of valid H-bonds between 267 and 292 (Supplementary Material Graphs SI152-SI159). The atom groups maintaining these hydrogen interactions were mainly the oxygen-phosphate groups (O1P and O2P) and oxygen's (O2, O3, O4, O5) as previously described for DYS19A (Supplementary Material Graphs SI214-SI241).

The observed H-bond pattern in DYS531 was nearly identical to the other Y-STRs (Supplementary Material Graphs SI160-SI169). The first water shell was detected between ≈2.8 Å and 3 Å distance. A few water molecules presented occupancy values near 100% during the MD simulations. The water interactions were mainly with oxygen-phosphate groups and oxygen's (excluding O2) when considering H-bond acceptors (Supplementary Material Graphs SI242-SI281).

The "spine of hydration" described in B-DNA minor groove molecular systems [37,39,76] is not observed in single-stranded molecular systems. The tested models do not show solvent peaks of hydration like the ones observed in minor groove double-helix DNA. Although there are no peaks of water molecules around these DNA strands, some specific water residues have high occupancy values during the MD simulation.

Molecular Dynamics: X3DNA and Curves+ Analysis

We have analysed the status of canonical and non-canonical base pairing with two or more H-bonds during the last 4 ns of UNAFold prediction tested models. The relative offset of the two base origins in the mean base pair plane is defined by Shear and Stretch, and the angle between the two x-axes considering the average normal to the base pair plane is the Opening [70]. We noticed that the Shear, Stretch and Opening critical parameters were significantly different (using as comparison the DNA double-helix Watson-Crick base pairs) [83,84], which implies that base pair geometry is different from double-helix DNA strands base pairing. The base pairs occurred in different number for each Y-STR (Table 5) and were not associated with any specific region which supports the putative formation of hairpins in different regions of the STR (Supplementary Material Table SI33). We have also noticed that the A<->T and T<->A (UNAFold predicted interaction) base pairing was not the only interaction detected. The A<->G, G<->G, T<->G, C<->A, T<->C, and A<->A base pairings were also present.

Molecular Dynamics: Same Length Y-STRs

The hairpin presence in DYS19A, DYS391 and DYS531 molecular systems with 11 repeats length were very similar when compared with the middle range tested alleles. Both RMSD and end to end distances shared a similar behaviour with the initial tested models. The small hairpins were maintained during the 12 ns MD simulation. There were some differences associated with the repeated motif of each STR (DYS19A:TAGA; DYS391:TCTA; DYS531:AAAT), mainly related with the stems localization. The overall results were in agreement with the previous reported results. The small hairpins present during the MD were, again, associated to electronegative pockets of Na+ that were present near hairpins stems or loops (<10 Å).

Discussion

Our results indicate that different forms of hairpins are present during MD simulations without a clear association to any specific initial constraints (considering all tested models), which suggests that formation of small hairpins can occur in any region of the STR. We also found that the occurrence of folded hairpin structures depends on specific conditions of the STR region (base composition, counterions, and water distribution). Although these results cannot be extrapolated to predict the STR regions with more probability to form hairpin structures, it can be concluded that hairpin structure(s) occurs often along STRs in short periods of time (ps). The loop formation of this type of hairpin can happen after fast folding trajectories following a downhill-like or direct folding [67,85]. This model can help to explain why strand slippage occurs during DNA replication. For the structures predicted by UNAFold we observed that, upon some conformation rearrangement, stabilization was achieved and maintained during 8 ns. The predicted small stem-loops located at 5' STR regions, when maintained during MD simulation, were very stable when compared with linear single-stranded conformations. This result suggests that small loops can occur in these repetitive regions even when UNAFold stability prediction is low (near zero free energy values). The occurrence of these small hairpins is related with the strand-slippage replication (Figure 1C) mechanism and can influence the polymerase enzyme during replication [22-25,77,86].

The H-bond base-pairing analyses revealed that the hairpin structures predicted by UNAFold can occur, but the conformational behaviour of different STRs is dependent of the repeat sequence and length of the STR. We also observed hairpin structures in the single-stranded systems, and that specific initial constraint (UNAFold predictions) does not influence significantly the putative occurrence of small hairpin structures. Strong base-pairing with correct geometric Watson-Crick interactions (canonical and non-canonical) was detected along each Y-STR between different nucleotides, which denotes that nucleotide interactions were not biased to A<->T and T<->A connections as predicted by UNAFold.

In previous works, MD of DNA (mainly in B-DNA helix) treated water as a dielectric continuum ignoring its capacity for hydration, bonding, and solvation in different modes [76]. Here we have provided an in-depth analysis of local and global counterion and water distribution. A relatively stable first water shell was observed in all MD simulations which contributed to the equilibrium of DNA single-stranded molecules with hairpins. The counterion distribution showed that electronegative pockets of Na+ were present near hairpins stems or loops (<10 Å), which corroborates their importance to the process of hairpin formation (stem base pairing occurrence followed by loop nucleation). There was specific folding potential associated to each Y-STR (small hairpins presence) which probably will impact STRs length variation resulting from replication errors (e.g., strand-slippage

replication mechanism, Figure 1C). Since we are using Y-STRs, the folding potential that produces the hairpin structures is definitively the main cause of strand-slippage replication errors and consequently allele size variation, because recombination does not occur in these regions [15,34]. Although MD simulations have limitations, the conformations in single-stranded and UNAFold predicted strands of DNA were consistent with previous studies [39-42,46,55,87,88]. The particular conditions tested *in silico* with these simulations, mimicking DNA systems that are very close to *in vitro* conditions in PCR, showed the importance of local environments to the conformational status of DNA strands that are being replicated.

Slipped-stranded DNA (S-DNA) conformation (hairpins structures or single-stranded loops) can occur in direct tandem repeats [89]. The repeated sequence of STRs can contribute to the mispairing of complementary repeats upon denaturing and renaturing (conditions simulated in the MDs of each STR). The release of DNA torsional stress can lead to formation of hydrogen bonds for some of the repeated units, leading to the occurrence of stable hairpins that result from loop nucleation. In these cases local DNA changes can lead to supercoiled relaxation because the supercoiling [90-92] is energetically unfavourable. We detected other conformations that were putatively associated with more stable hairpin structures, and characterized by a positive supercoiling. These positive supercoiled structures will tend to overtwist [89,92]. Different patterns were observed in the three different Y-STRs, and interchanging states between positive and negative supercoiled conformations, that might have implications in the influence of supercoiling in transcription [93,94] and replication [95] of STR genomic regions. The different types of torsion and tension observed in the DNA models tested may facilitate or block the action of topoisomerases and polymerases during replication, inducing errors like strand-slippage.

The variation in STRs length may have impact in the genome architecture by altering the state of the chromatin conformation or by promoting the insertion of transposable elements [33]. Since STRs are widely spread throughout human genome and are involved in transcription and signalling pathways, our results can be used to analyse STR instability in other regions of the genome that are associated with disease (e.g., cancer, neurodegenerative disorders) [18,80,96,97]. Testing the behaviour of STRs is of great importance to understand instability of breakpoint regions linked to human disease. The model here proposed for the formation of small/medium hairpins in Y-STRs of four nucleotide motifs can easily be applied to understand the critical point that separates single step mutation patterns from large expansion occurring in some STRs regions [23,79,80,97-99], and even the binding properties of some specific repetitive sequences. Ultimately, these results may have implications in understanding neurodegenerative disorders and may as well explain why genomic repetitive segments are involved in these types of disease. These results helped to understand the evolutionary dynamics of these genomic regions, and can

be used to study the processes behind generation of disease-related expansions. Coding and non-coding STRs share mechanisms that can be used to explain the main features of human disease that result from STR expansions [100]. The process of largest expansion can result from the formation of largest hairpin structures and/or compact conformational states with high free energetic values in both non-dividing and dividing cells.

Supplementary Data are available online: Supplementary Tables SI1-SI33, Supplementary Figures SI1-SI2, and Supplementary Graphs SI1-SI281. NABpy Python Supplementary File.

Funding

This work was partially supported by research grants to JC from the Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP) and FP (SFRH/BPD/44637/2008) from the Portuguese Foundation for Science and Technology (FCT). IPATIMUP is an Associate Laboratory of the Portuguese Ministry of Science, Technology and Higher Education and is partially supported by FCT.

References

1. Lynch M (2007) The origins of genome architecture. The University of Michigan: Sinauer Associates, 2007. 494 p.
2. Locey KJ, White EP (2011) Simple structural differences between coding and noncoding DNA. PLoS One 6: e14651.
3. Greenbaum JA, Pang B, Tullius TD (2007) Construction of a genome-scale structural map at single-nucleotide resolution. Genome Res 17: 947-953.
4. Consortium EP, Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799-816.
5. Sastre L (2012) Clinical implications of the ENCODE project. Clin Transl Oncol 14: 801-802.
6. Consortium EP (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306: 636-640.
7. de Souza N (2012) The ENCODE project. Nat Methods 9: 1046.
8. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, et al. (2012) GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res 22: 1760-1774.
9. Pennisi E (2012) Genomics. ENCODE project writes eulogy for junk DNA. Science 337: 1159, 1161.
10. Farnham PJ (2012) Thematic minireview series on results from the ENCODE Project: Integrative global analyses of regulatory regions in the human genome. J Biol Chem 287: 30885-30887.
11. Ballantyne KN, Goedbloed M, Fang R, Schaap O, Lao O, et al. (2010) Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. Am J Hum Genet 87: 341-353.
12. Zhivotovsky LA, Underhill PA, Cinnioglu C, Kayser M, Morar B, et al. (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet 74: 50-61.
13. Jobling MA (2004) Human Evolutionary Genetics: Origins, Peoples & Disease; Jobling MAea, editor. New York, USA.: Garland Science.
14. Jarve M, Zhivotovsky LA, Rootsi S, Help H, Rogaev EI, et al. (2009) Decreased rate of evolution in Y chromosome STR loci of increased size of the repeat unit. PLoS One 4: e7276.
15. Kayser M, Caglia A, Corach D, Fretwell N, Gehrig C, et al. (1997) Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med 110: 125-133, 141-129.
16. Kayser M, Sajantila A (2001) Mutations at Y-STR loci: implications for paternity testing and forensic analysis. Forensic Sci Int 118: 116-121.
17. Gusmao L, Krawczak M, Sanchez-Diz P, Alves C, Lopes A, et al. (2003) Bimodal allele frequency distribution at Y-STR loci DYS392 and DYS438: no evidence for a deviation from the stepwise mutation model. Int J Legal Med 117: 287-290.
18. Madsen BE, Villesen P, Wiuf C (2008) Short tandem repeats in human exons: a target for disease mutations. BMC Genomics 9: 410.
19. Fan H, Chu JY (2007) A brief review of short tandem repeat mutation. Genomics Proteomics Bioinformatics 5: 7-14.
20. Thomas DC, Roberts JD, Fitzgerald MP, Kunkel TA (1990) Fidelity of animal cell DNA polymerases alpha and delta and of a human DNA replication complex. Basic Life Sci 52: 289-297.
21. Thomas DC, Roberts JD, Sabatino RD, Myers TW, Tan CK, et al. (1991) Fidelity of mammalian DNA replication and replicative DNA polymerases. Biochemistry 30: 11751-11759.
22. Samadashwily GM, Raca G, Mirkin SM (1997) Trinucleotide repeats affect DNA replication *in vivo*. Nat Genet 17: 298-304.

23. Pelletier R, Krasilnikova MM, Samadashwily GM, Lahue R, Mirkin SM (2003) Replication and expansion of trinucleotide repeats in yeast. Mol Cell Biol 23: 1349-1357.
24. Krasilnikova MM, Mirkin SM (2004) Replication stalling at Friedreich's ataxia (GAA)n repeats *in vivo*. Mol Cell Biol 24: 2286-2295.
25. Chandok GS, Patel MP, Mirkin SM, Krasilnikova MM (2012) Effects of Friedreich's ataxia GAA repeats on DNA replication in mammalian cells. Nucleic Acids Res 40: 3964-3974.
26. Ruitberg CM, Reeder DJ, Butler JM (2001) STRBase: a short tandem repeat DNA database for the human identity testing community. Nucleic Acids Res 29: 320-322.
27. Butler JM, Decker AE, Kline MC, Vallone PM (2005) Chromosomal duplications along the Y-chromosome and their potential impact on Y-STR interpretation. J Forensic Sci 50: 853-859.
28. Gusmao L, Alves C (2005) Y chromosome STR typing. Methods Mol Biol 297: 67-82.
29. Decker AE, Kline MC, Redman JW, Reid TM, Butler JM (2008) Analysis of mutations in father-son pairs with 17 Y-STR loci. Forensic Sci Int Genet 2: e31-35.
30. Balaresque P, Parkin EJ, Roewer L, Carvalho-Silva DR, Mitchell RJ, et al. (2009) Genomic complexity of the Y-STR DYS19: inversions, deletions and founder lineages carrying duplications. Int J Legal Med 123: 15-23.
31. Voineagu I, Freudenreich CH, Mirkin SM (2009) Checkpoint responses to unusual structures formed by DNA repeats. Mol Carcinog 48: 309-318.
32. Shi W, Ayub Q, Vermeulen M, Shao RG, Zuniga S, et al. (2010) A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations. Mol Biol Evol 27: 385-393.
33. Farre M, Bosch M, Lopez-Giraldez F, Ponsa M, Ruiz-Herrera A (2011) Assessing the role of tandem repeats in shaping the genomic architecture of great apes. PLoS One 6: e27239.
34. Gusmao L, Sanchez-Diz P, Calafell F, Martin P, Alonso CA, et al. (2005) Mutation rates at Y chromosome specific microsatellites. Hum Mutat 26: 520-528.
35. Vicard P, Dawid AP, Mortera J, Lauritzen SL (2008) Estimating mutation rates from paternity casework. Forensic Sci Int Genet 2: 9-18.
36. Vicard P, Dawid AP (2004) A statistical treatment of biases affecting the estimation of mutation rates. Mutat Res 547: 19-33.
37. Young MA, Ravishanker G, Beveridge DL (1997) A 5-nanosecond molecular dynamics trajectory for B-DNA: analysis of structure, motions, and solvation. Biophys J 73: 2313-2336.
38. Sarkar A, Eroglu S, Poirier MG, Gupta P, Nemani A, et al. (2002) Dynamics of chromosome compaction during mitosis. Exp Cell Res 277: 48-56.
39. Chuprina VP, Heinemann U, Nurislamov AA, Zielenkiewicz P, Dickerson RE, et al. (1991) Molecular dynamics simulation of the hydration shell of a B-DNA decamer reveals two main types of minor-groove hydration depending on groove width. Proc Natl Acad Sci U S A 88: 593-597.
40. Beveridge DL, Barreiro G, Byun KS, Case DA, Cheatham TE, 3rd, et al. (2004) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps. Biophys J 87: 3799-3813.
41. Portella G, Orozco M (2010) Multiple routes to characterize the folding of a small DNA hairpin. Angew Chem Int Ed Engl 49: 7673-7676.
42. Sagui C, Darden TA (1999) Molecular dynamics simulations of biomolecules: long-range electrostatic effects. Annu Rev Biophys Biomol Struct 28: 155-179.
43. Lei H, Duan Y (2008) Protein folding and unfolding by all-atom molecular dynamics simulations. Methods Mol Biol 443: 277-295.
44. Brooks CL, 3rd (1998) Simulations of protein folding and unfolding. Curr Opin Struct Biol 8: 222-226.

45. Meller J (2001) Molecular Dynamics. ENCYCLOPEDIA OF LIFE SCIENCES. Cornell University, Ithaca, New York, USA Nicholas Copernicus University: Nature Publishing Group.
46. Auffinger P, Westhof E (1998) Simulations of the molecular dynamics of nucleic acids. Curr Opin Struct Biol 8: 227-236.
47. Markham NR, Zuker M (2008) UNAFold: software for nucleic acid folding and hybridization. Methods Mol Biol 453: 3-31.
48. Damas J, Carneiro J, Goncalves J, Stewart JB, Samuels DC, et al. (2012) Mitochondrial DNA deletions are associated with non-B DNA conformations. Nucleic Acids Res.
49. Brazdova M, Navratilova L, Tichy V, Nemcova K, Lexa M, et al. (2013) Preferential binding of hot spot mutant p53 proteins to supercoiled DNA *in vitro* and in cells. PLoS One 8: e59567.
50. Pereira F, Soares P, Carneiro J, Pereira L, Richards MB, et al. (2008) Evidence for variable selective pressures at a large secondary structure of the human mitochondrial DNA control region. Mol Biol Evol 25: 2759-2770.
51. SantaLucia J, Jr., Hicks D (2004) The thermodynamics of DNA structural motifs. Annu Rev Biophys Biomol Struct 33: 415-440.
52. Jones RM, Petermann E (2012) Replication fork dynamics and the DNA damage response. Biochem J 443: 13-26.
53. Kar RK, Suryadevara P, Jana J, Bhunia A, Chatterjee S (2012) Novel G-quadruplex stabilizing agents: in-silico approach and dynamics. J Biomol Struct Dyn Epub: 1–22.
54. Robbins TJ, Wang Y (2012) Effect of initial ion positions on the interactions of monovalent and divalent ions with a DNA duplex as revealed with atomistic molecular dynamics simulations. J Biomol Struct Dyn Epub: 1-13.
55. Mukherjee S, Bhattacharyya D (2012) Influence of divalent magnesium ion on DNA: molecular dynamics simulation studies. J Biomol Struct Dyn Epub: 1-17.
56. Berman HM, Westbrook J, Feng Z, Iype L, Schneider B, et al. (2003) The nucleic acid database. Methods Biochem Anal 44: 199-216.
57. Rychlik W, Spencer WJ, Rhoads RE (1990) Optimization of the annealing temperature for DNA amplification *in vitro*. Nucleic Acids Res 18: 6409-6412.
58. Kramer MF, Coen DM (2006) Enzymatic amplification of DNA by PCR: standard procedures and optimization. Curr Protoc Cytom Appendix 3: Appendix 3K.
59. D.A. Case TAD, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A.W. Goetz, I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, R.M. Wolf, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P.A. Kollman (2012) AMBER 12. University of California, San Francisco.
60. Cheatham TE, 3rd, Cieplak P, Kollman PA (1999) A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. J Biomol Struct Dyn 16: 845-862.
61. Jackson RM, Gabb HA, Sternberg MJE (1998) Rapid refinement of protein interfaces incorporating solvation: Application to the docking problem. Journal of molecular biology 276: 265-285.
62. Loncharich RJ, Brooks BR, Pastor RW (1992) Langevin dynamics of peptides- the frictional dependence of isomerization rates of n-acetylalanyl-n-methylamide Biopolymers 32: 523-535.
63. Darden T, York D, Pedersen L (1993) Particle mesh Ewald- an n.log(n) method for ewald sums in large systems Journal of Chemical Physics 98: 10089-10092.
64. Ryckaert JP, Ciccotti G, Berendsen HJC (1977) Numerical integration of cartesian equations of motion of a system with constraints- molecular dynamics of n-alkanes. Journal of Computational Physics 23: 327-341.

65. Case DA, Cheatham TE, 3rd, Darden T, Gohlke H, Luo R, et al. (2005) The Amber biomolecular simulation programs. J Comput Chem 26: 1668-1688.
66. Baker ES, Dupuis NF, Bowers MT (2009) DNA hairpin, pseudoknot, and cruciform stability in a solvent-free environment. J Phys Chem B 113: 1722-1727.
67. Ma HT, On KF, Tsang YH, Poon RY (2007) An inducible system for expression and validation of the specificity of short hairpin RNA in mammalian cells. Nucleic Acids Res 35: e22.
68. Goddard NL, Bonnet G, Krichevsky O, Libchaber A (2000) Sequence dependent rigidity of single stranded DNA. Phys Rev Lett 85: 2400-2403.
69. Software StatSoft I (2011) STATISTICA (data analysis software system). version 10 ed.
70. Lu XJ, Olson WK (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res 31: 5108-5121.
71. Lavery R, Moakher M, Maddocks JH, Petkeviciute D, Zakrzewska K (2009) Conformational analysis of nucleic acids revisited: Curves+. Nucleic Acids Res 37: 5917-5929.
72. Perez A, Marchan I, Svozil D, Sponer J, Cheatham TE, 3rd, et al. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. Biophys J 92: 3817-3829.
73. Sponer J, Jurecka P, Hobza P (2004) Accurate interaction energies of hydrogen-bonded nucleic acid base pairs. J Am Chem Soc 126: 10142-10151.
74. Kabelac M, Kratochvil M, Sponer J, Hobza P (2000) Structure, energetics, vibrational frequencies and charge transfer of base pairs, nucleoside pairs, nucleotide pairs and B-DNA pairs of trinucleotides: ab initio HF/MINI-1 and empirical force field study. J Biomol Struct Dyn 17: 1077-1086.
75. Jurecka P, Hobza P (2003) True stabilization energies for the optimal planar hydrogen-bonded and stacked structures of guanine...cytosine, adenine...thymine, and their 9- and 1-methyl derivatives: complete basis set calculations at the MP2 and CCSD(T) levels and comparison with experiment. J Am Chem Soc 125: 15608-15613.
76. Jayaram B, Beyeridge DL (1996) Modeling DNA in aqueous solutions: theoretical and computer simulation studies on the ion atmosphere of DNA. Annu Rev Biophys Biomol Struct 25: 367-394.
77. Kunkel TA (2009) Evolving views of DNA replication (in)fidelity. Cold Spring Harb Symp Quant Biol 74: 91-101.
78. Kunkel TA (2004) DNA replication fidelity. J Biol Chem 279: 16895-16898.
79. Gatchel JR, Zoghbi HY (2005) Diseases of unstable repeat expansion: mechanisms and common principles. Nat Rev Genet 6: 743-755.
80. Usdin K (2008) The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. Genome Res 18: 1011-1019.
81. Marvin DA, Spencer M, Wilkins MH, Hamilton LD (1958) A new configuration of deoxyribonucleic acid. Nature 182: 387-388.
82. Pohl FM, Jovin TM (1972) Salt-induced co-operative conformational change of a synthetic DNA: equilibrium and kinetic studies with poly (dG-dC). J Mol Biol 67: 375-396.
83. Lu XJ, Olson WK (2008) 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nat Protoc 3: 1213-1227.
84. Zheng G, Lu XJ, Olson WK (2009) Web 3DNA--a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. Nucleic Acids Res 37: W240-246.
85. Ma H, Wan C, Wu A, Zewail AH (2007) DNA folding and melting observed in real time redefine the energy landscape. Proc Natl Acad Sci U S A 104: 712-716.
86. Canceill D, Viguera E, Ehrlich SD (1999) Replication slippage of different DNA polymerases is inversely related to their strand displacement efficiency. J Biol Chem 274: 27481-27490.

87. Dixit SB, Beveridge DL, Case DA, Cheatham TE, 3rd, Giudice E, et al. (2005) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. Biophys J 89: 3721-3740.
88. Goh GB, Knight JL, Brooks CL, 3rd (2012) Constant pH Molecular Dynamics Simulations of Nucleic Acids in Explicit Solvent. J Chem Theory Comput 8: 36-46.
89. Mirkin SM (2001) DNA Topology: Fundamentals ENCYCLOPEDIA OF LIFE SCIENCES. University of Illinois at Chicago, Illinois, USA: Nature Publishing Group.
90. Mirkin SM, Bogdanova ES, Gorlenko Ah M, Gragerov AI, Larionov OA (1979) [Effect of DNA supercoiling on transcription performed by normal and mutant Escherichia coli RNA-polymerases]. Mol Biol (Mosk) 13: 1341-1349.
91. Mirkin SM, Bogdanova ES, Gorlenko ZM, Gragerov AI, Larionov OA (1979) DNA supercoiling and transcription in Escherichia coli: influence of RNA polymerase mutations. Mol Gen Genet 177: 169-175.
92. Krasilnikov AS, Podtelezhnikov A, Vologodskii A, Mirkin SM (1999) Large-scale effects of transcriptional DNA supercoiling *in vivo*. J Mol Biol 292: 1149-1160.
93. Droge P (1993) Transcription-driven site-specific DNA recombination *in vitro*. Proc Natl Acad Sci U S A 90: 2759-2763.
94. Margolin P, Zumstein L, Sternglanz R, Wang JC (1985) The Escherichia coli supX locus is topA, the structural gene for DNA topoisomerase I. Proc Natl Acad Sci U S A 82: 5437-5441.
95. Witz G, Stasiak A (2010) DNA supercoiling and its role in DNA decatenation and unknotting. Nucleic Acids Res 38: 2119-2133.
96. Whitehouse I, Owen-Hughes T (2010) ATRX: Put me on repeat. Cell 143: 335-336.
97. Mirkin SM (2006) DNA structures, repeat expansions and human hereditary disorders. Curr Opin Struct Biol 16: 351-358.
98. Jochens A, Caliebe A, Rosler U, Krawczak M (2011) Empirical evaluation reveals best fit of a logistic mutation model for human Y-chromosomal microsatellites. Genetics 189: 1403-1411.
99. Siianova E, Mirkin SM (2001) [Expansion of trinucleotide repeats]. Mol Biol (Mosk) 35: 208-223.
100. McMurray CT (2010) Mechanisms of trinucleotide repeat instability during human development. Nat Rev Genet 11: 786-799.

Tables

Table 1: Accuracy results of UNAFold to predict hairpin structures using a database of experimentally determined NMR structures of 22 single-stranded DNA (ssDNA) with hairpins.

PDB ID	3D structure type	UNAFold prediction (base pairing and loop)
1AC7	SsDNA hairpin	Identical to 3D
1BJH	SsDNA hairpin	Identical to 3D
1DGO	SsDNA hairpin	Identical to 3D
1ECU	SsDNA hairpin	Identical to 3D
1EL2	DNA complex telomeric structure (3 loops)	Different from 3D
1ELN	DNA complex telomeric structure (3 loops)	Different from 3D
1EN1	Primer binding site HIV -ssDNA hairpin	Identical to 3D
1FV8	SsDNA heterochiral hairpin	Identical to 3D
1IDX	SsDNA hairpin	Identical to 3D
1II1	SsDNA hairpin	Identical to 3D
1JVE	AT-Rich ssDNA with the GAA-Hairpin Loop	Identical to 3D
1KR8	SsDNA hairpin	Identical to 3D
1LA8	SsDNA hairpin	Identical to 3D
1LAE	SsDNA hairpin	Identical to 3D
1NGO	SsDNA hairpin	Identical to 3D
1NGU	SsDNA hairpin	Identical to 3D (except 2 residues in stem)
1P0U	SsDNA triplet repeats hairpin	Identical to 3D
1PQT	SsDNA hairpin	Identical to 3D
1QE7	SsDNA hairpin	Identical to 3D
1XUE	SsDNA hairpin	Identical to 3D only for loop
1ZHU	Hairpin loop formed by the DNA triplet GCA.	Identical to 3D
2M22	SsDNA hairpin	Identical to 3D

Table 2: Sequence/structural data analysed in UNAFold for 3 tetranucleotide STRs of NIST database. The STR sequence is represented as (repeat motif) number of repeats. Connected bases correspond to stem connected nucleotides in hairpin structure.

	Motif length	STR Sequence	UNAFold prediction	Number of loops	Stem lenght (bp)	Loop size or Bulge	Free energy (Kcal/mol)	Connected bases
DYS19A	4	(TAGA)3(TAGG)1(TAGA)11	YES	1	2	6	1.57	A<->T;T<->A
DYS391	4	(TCTA)10	YES	1	2	4	1.44	A<->T;T<->A
DYS531	4	(AAAT)11	YES	1	2	6	1.65	A<->T;T<->A

Table 3: H-bonds present during last 4 ns of molecular dynamics of UNAFold predicted stem region for DYS19A; Acceptor and donor represent hydrogen bond acceptor and donor respectively.

Residue number acceptor	Atom acceptor	Residue number donor	Hydrogen Atom donor	Atom donor	%occupied
4	N7	12	H62	N6	8.68
5	O4	6	H61	N6	28.14
12	N7	5	H3	N3	20.96
6	N1	5	H3	N3	14.97
5	O4	6	H62	N6	12.87
12	N3	5	H3	N3	1.5
12	N1	5	H3	N3	0.3
5	O4	12	H61	N6	0.3
6	N3	11	H22	N2	55.09
6	N1	11	H22	N2	21.26
7	O6	6	H62	N6	20.66
6	N1	7	H1	N1	2.69
6	N1	7	H21	N2	1.5
6	N3	7	H22	N2	0.3
6	N7	11	H22	N2	0.3

Table 4: Single-stranded H-bonds descriptive statistics of single-stranded Y-STR DYS19A molecular dynamics simulation from 8 to 12 ns; UNAFold prediction H-bond global statistics of Y-STR DYS19A molecular dynamics simulation with UNAFold predicted constraints from 8 to 12 ns; H-bond global statistics of Y-STR DYS19A molecular dynamics simulation with UNAFold predicted constraints in the end of the DNA molecule (stem with base pair between 48<->57 and 49<->56) from 8 to 12 ns. Described values of H-bond percentage of occupancy during molecular dynamics (%occupied), distance (Å) of acceptor-donor H-bond (distance), angle of H-bond (angle), lifetime and maxocc as calculated by ptraj hbond function in the AMBER package [59].

		Valid H-Bonds	Mean	Minimum	Maximum	Std.Dev.
Single-stranded	**%occupied**	182	12.38	0.30	97.01	22.26
	distance	182	4.24	2.92	4.98	0.57
	angle	182	47.21	20.74	59.82	9.61
	lifetime	182	11.56	4.00	190.70	24.73
	maxocc	182	8.87	1.00	133.00	21.17
UNAFold prediction	**%occupied**	239	8.89	0.30	99.70	18.02
	distance	239	4.21	2.87	5.00	0.56
	angle	239	47.41	16.97	59.99	9.08
	lifetime	239	11.07	4.00	666.00	44.25
	maxocc	239	8.16	1.00	239.00	24.32
UNAFold prediction (stem with base pair between 48<->57 and 49<->56)	**%occupied**	208	8.21	0.30	99.40	16.66
	distance	208	4.18	2.85	4.98	0.53
	angle	208	47.69	13.89	59.95	9.92
	lifetime	208	10.61	4.00	442.70	34.87
	maxocc	208	7.04	1.00	144.00	20.87

Table 5: Local base-pair and helical parameters of DYS19A, DYS391 and DYS531 UNAFold predicted molecular systems (average PDB of last 4ns), as calculated by X3DNA software.

Model	Nucleotides	Shear	Stretch	Stagger	Buckle	Prop-Tw	Opening	X-disp	Y-disp	h-Rise	Incl.	Tip	h-Twist
DYS19A	A+G (2+15)	-3.19	-7.42	1.31	24.73	-26.21	-170.69	0	0	0	0	0	0
UNAFold	G+G (7+11)	3.62	7.41	-0.77	-32.3	-36.06	179.02	-5.86	-6.84	-1.11	-29.35	19.02	157.87
prediction	T+G (13+19)	-4.88	-0.7	-0.07	55.94	32.28	64.22	1.61	5.63	7.76	72.77	-16.81	-162.84
	T-A (17-32)	-2.84	2.68	0.9	21.41	14.52	-99.59	-0.26	-7.88	-1.19	43.53	12.65	-171.64
	T+G (21+31)	0.12	-5.72	-0.43	-5.84	-20.15	-132.29	-1.81	-0.21	-3.6	1.87	-3.53	135.85
	A+T (48+53)	0.92	-3.35	-1.33	22.92	-3.64	68.96	30.35	0.14	12.35	37.73	-61.13	-83.71
DYS391	C-A (30-32)	-1.29	-6.03	0.44	-28.93	24.67	154.87	0	0	0	0	0	0
UNAFold	C-A (34-36)	-2.86	-1.32	-2.27	-3.57	-17.22	174.98	4.6	-3.6	-6.02	-42.22	53.85	-151.76
prediction	T-C (37-38)	3.08	4.99	-1.43	4.65	37.86	141.64	6.94	2.32	3.3	65.77	-22.58	-136.99
DYS531	A-A (6-7)	2.3	4.91	0.65	-40.09	28.03	150.52	0	0	0	0	0	0
UNAFold	A-A (11-13)	-5.4	0.21	-1.35	40.93	14.66	43.65	3.57	-5.62	2.62	60.19	-14.96	159.35
prediction													

Figures

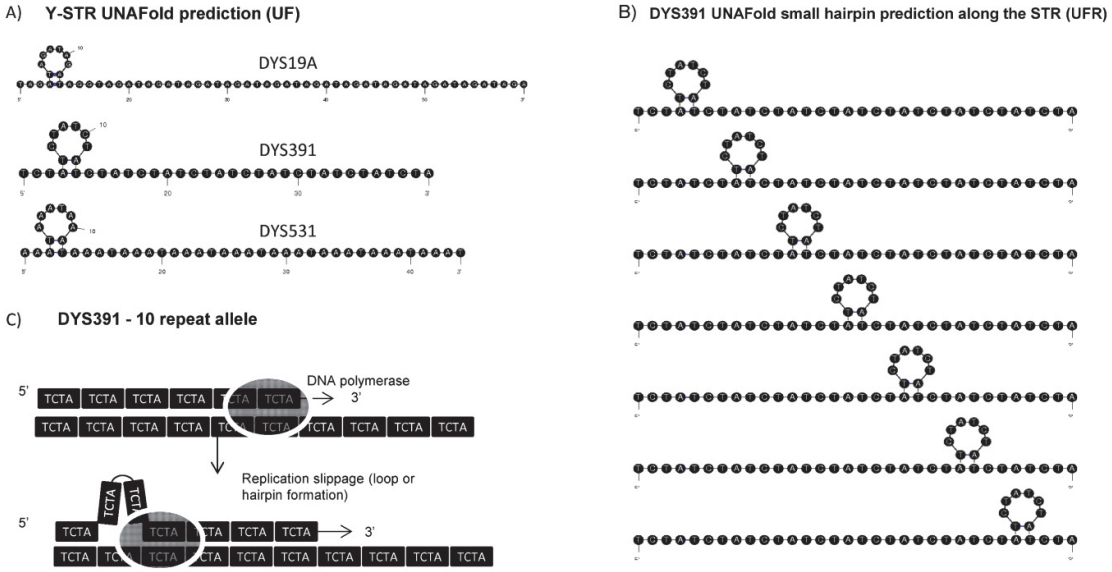


Figure 1: A) UNAFold predicted conformations (UF) of DYS19A (15 repeats), DYS391 (10 repeats) and DYS531 (11 repeats); B) Hairpin conformations derived from UNAFold loop prediction (UF) in different regions (UFR) of the DYS391 Y-STR; C) Example of strand-slippage replication mechanism occurring in DYS391 10 repeat allele.

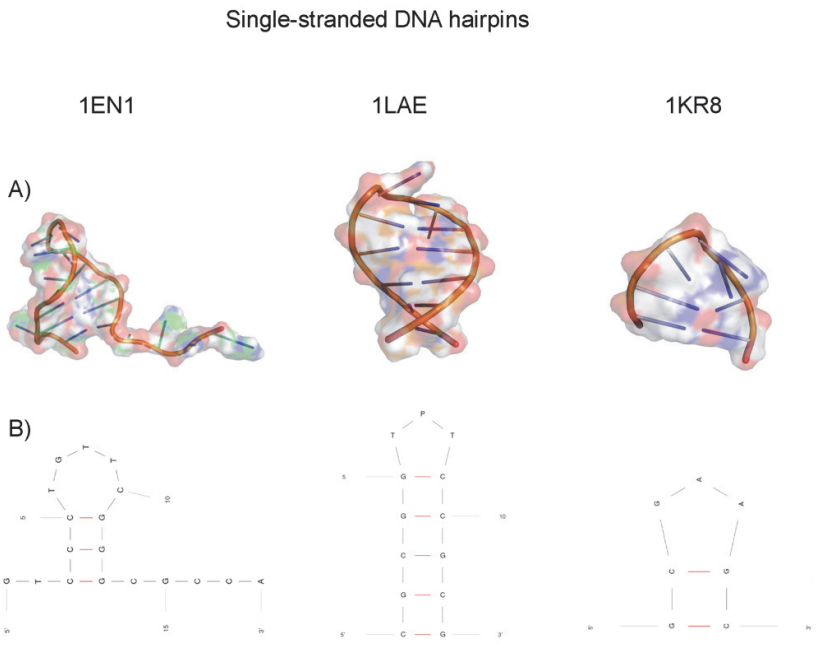


Figure 2: Molecular representation of single-stranded DNA hairpins for PDB ID 1EN1, 1LAE, and 1KR8 considering: A) 3D NMR single-stranded structures with small hairpins; B) UNAFold predictions using PDB sequence.

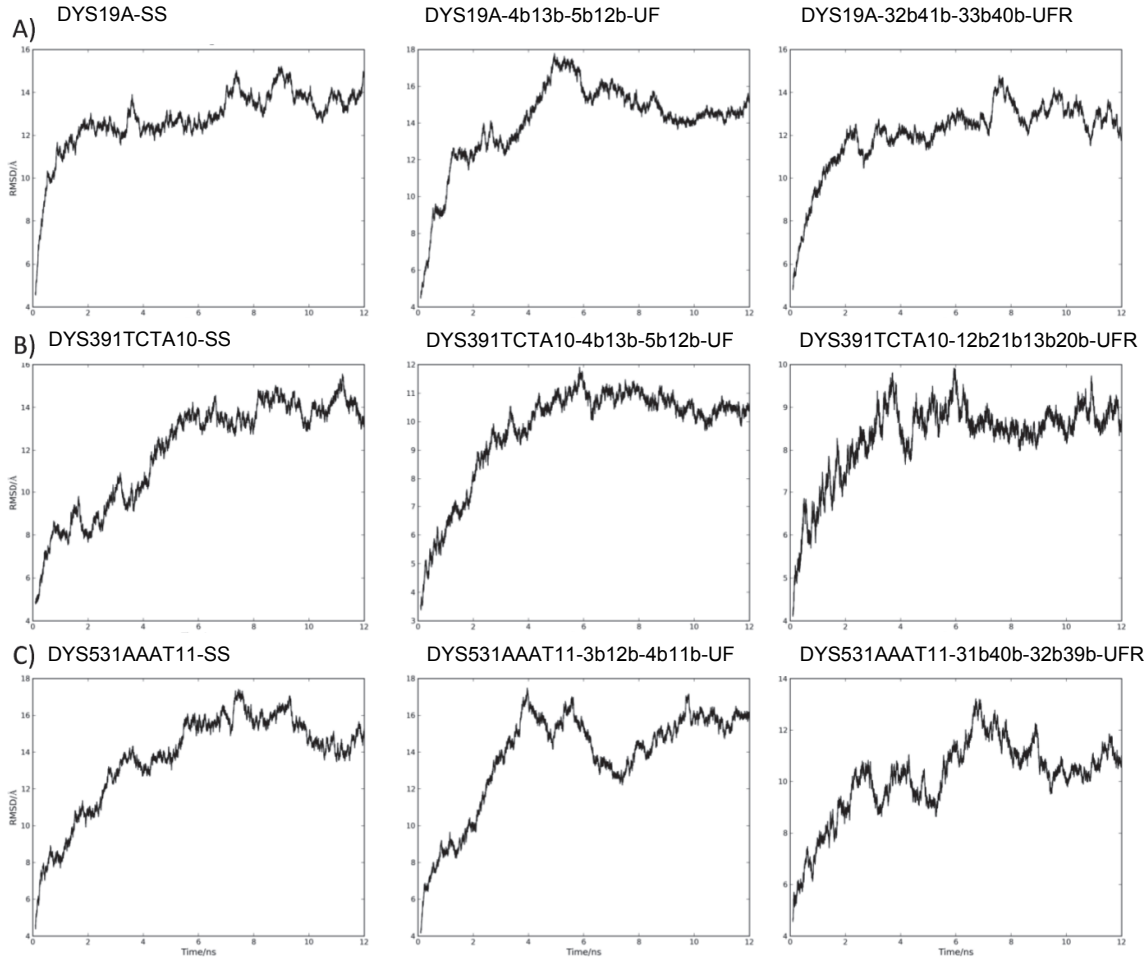


Figure 3: Root mean-square deviation values for backbone atoms: A) DYS19A single-stranded DNA molecule (left), DYS19A UNAFold prediction (middle), and DYS19A UNAFold predicted structure (identical loop with same base pair connections in stem) occupying different regions of the STR (right); B) DYS391 single-stranded DNA molecule (left), DYS391 UNAFold prediction (middle), and DYS391 UNAFold prediction in other region of STR (right); C) DYS531 single-stranded DNA molecule (left), DYS531 UNAFold prediction (middle), and DYS531 UNAFold prediction in other region of STR (right).

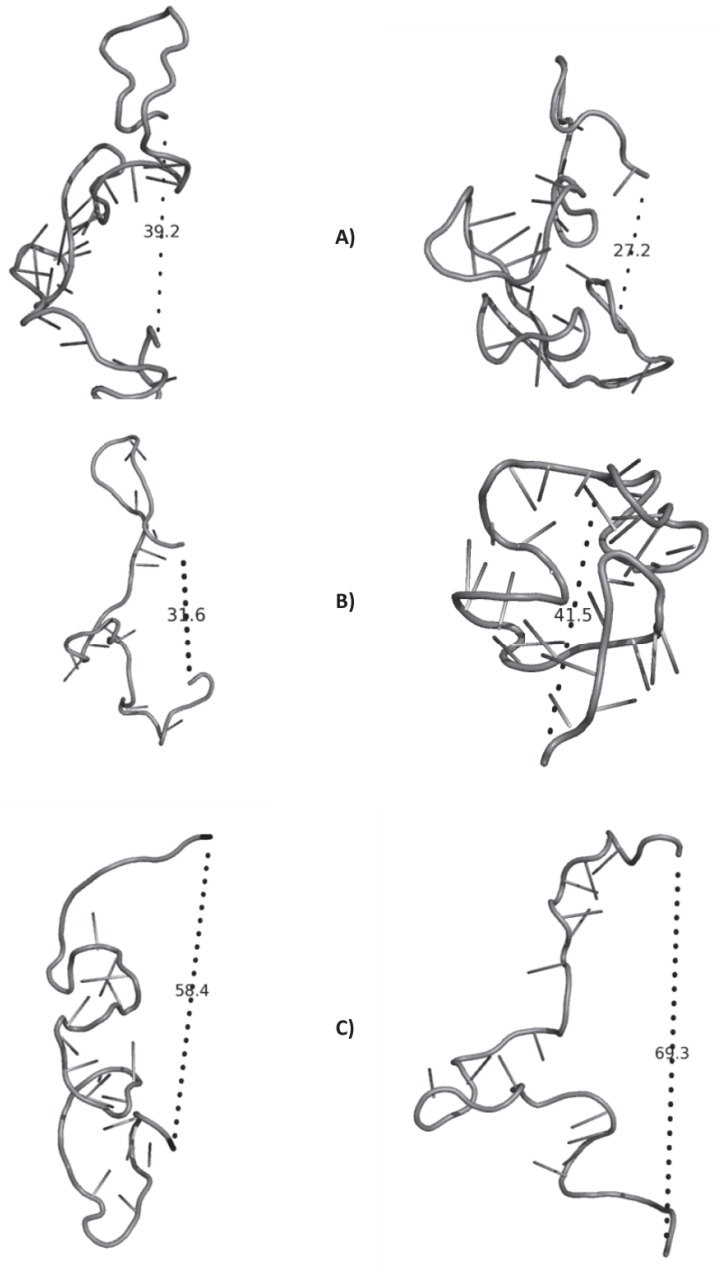


Figure 4: First residue carbon (C1') to last residue carbon (C1') distances in Å of a representative snapshot of DYS19A (A), DYS391 (B), and DYS531 (C). Single-stranded molecule is represented in left and UNAFold predicted molecule in right.

	H-bonds DYS391TCTA10-SS_min				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	147	5.20	0.30	77.54	10.30
distance	147	4.20	2.92	4.99	0.54
angle	147	45.81	12.81	59.94	10.56
lifetime	147	9.26	4.00	85.70	12.28
maxocc	147	6.22	1.00	114.00	14.41

	H-bonds DYS391TCTA10-4b13b-5b12b-UF				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	77	7.05	0.30	80.84	15.52
distance	77	4.27	2.94	4.97	0.50
angle	77	46.63	15.88	59.50	9.31
lifetime	77	6.96	4.00	45.70	7.19
maxocc	77	5.18	1.00	77.00	11.47

	H-bonds DYS391TCTA10-8b17b-9b16b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	123	7.77	0.30	97.31	15.36
distance	123	4.21	2.80	5.00	0.57
angle	123	46.81	18.47	59.73	9.93
lifetime	123	9.29	4.00	162.50	18.19
maxocc	123	8.44	1.00	194.00	24.90

	H-bonds DYS391TCTA10-12b21b-13b20b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	132	10.80	0.30	87.43	18.57
distance	132	4.19	3.00	5.00	0.56
angle	132	45.82	19.08	59.15	9.93
lifetime	132	11.66	4.00	149.30	19.99
maxocc	132	12.06	1.00	166.00	28.37

	H-bonds DYS391TCTA10-16b25b-17b24b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	144	7.12	0.30	97.31	15.79
distance	144	4.21	2.83	4.98	0.50
angle	144	47.60	14.35	59.79	9.09
lifetime	144	9.50	4.00	325.00	28.54
maxocc	144	5.80	1.00	166.00	16.94

	H-bonds DYS391TCTA10-20b29b-21b28b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	143	8.19	0.30	98.80	17.75
distance	143	4.19	2.96	4.97	0.53
angle	143	47.27	19.24	59.64	9.30
lifetime	143	10.81	4.00	330.00	30.50
maxocc	143	8.58	1.00	174.00	23.96

	H-bonds DYS391TCTA10-24b33b-25b32b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	89	8.90	0.30	97.90	19.03
distance	89	4.16	2.88	4.98	0.53
angle	89	47.78	22.79	59.68	8.60
lifetime	89	8.02	4.00	163.50	17.71
maxocc	89	6.06	1.00	188.00	20.94

	H-bonds DYS391TCTA10-28b37b-29b36b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	144	13.35	0.30	100.00	21.5
distance	144	4.27	2.89	4.99	0.6
angle	144	45.68	10.49	59.94	10.6
lifetime	144	29.17	4.00	1336.00	157.6
maxocc	144	14.56	1.00	334.00	46.5

Figure 5: Descriptive statistics of H-bonds (base pairing) in tested models of DY391 Y-STR. Described values of H-bond percentage of occupancy during molecular dynamics (%occupied), distance (Å) of acceptor-donor H-bond (distance), angle of H-bond (angle), lifetime and maxocc.

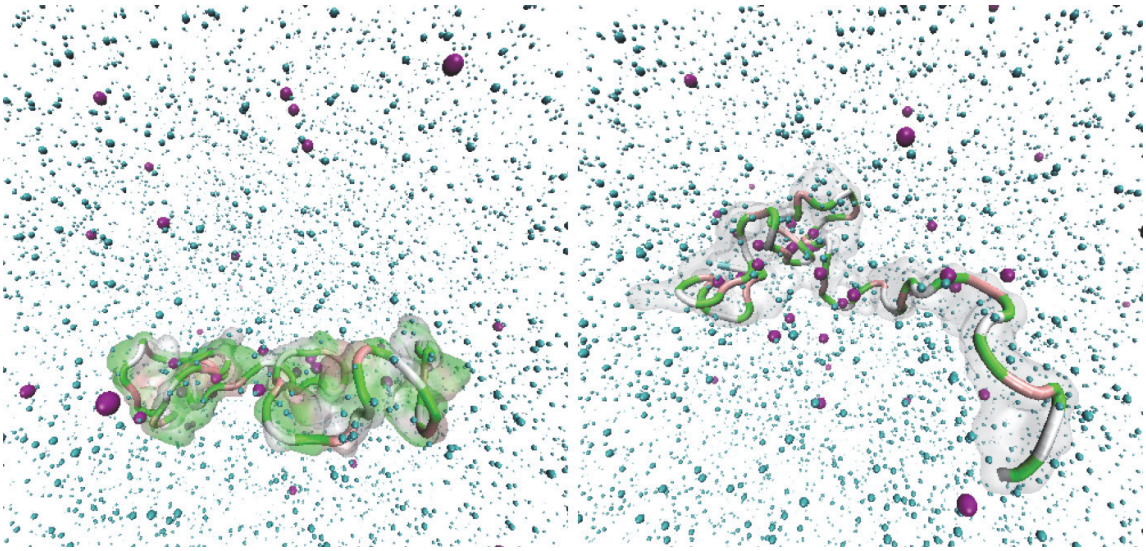


Figure 6: Supercoiled conformations of DYS19A-44b53b-45b52b-UFR (left) and DYS19A-48b57b-49b56b-UFR (right) complexes that are associated with high free energies and putatively more stable conformations [DT-light pink, DA- green, DG-white, Solvent (blue) and Na+ counterion (purple)].

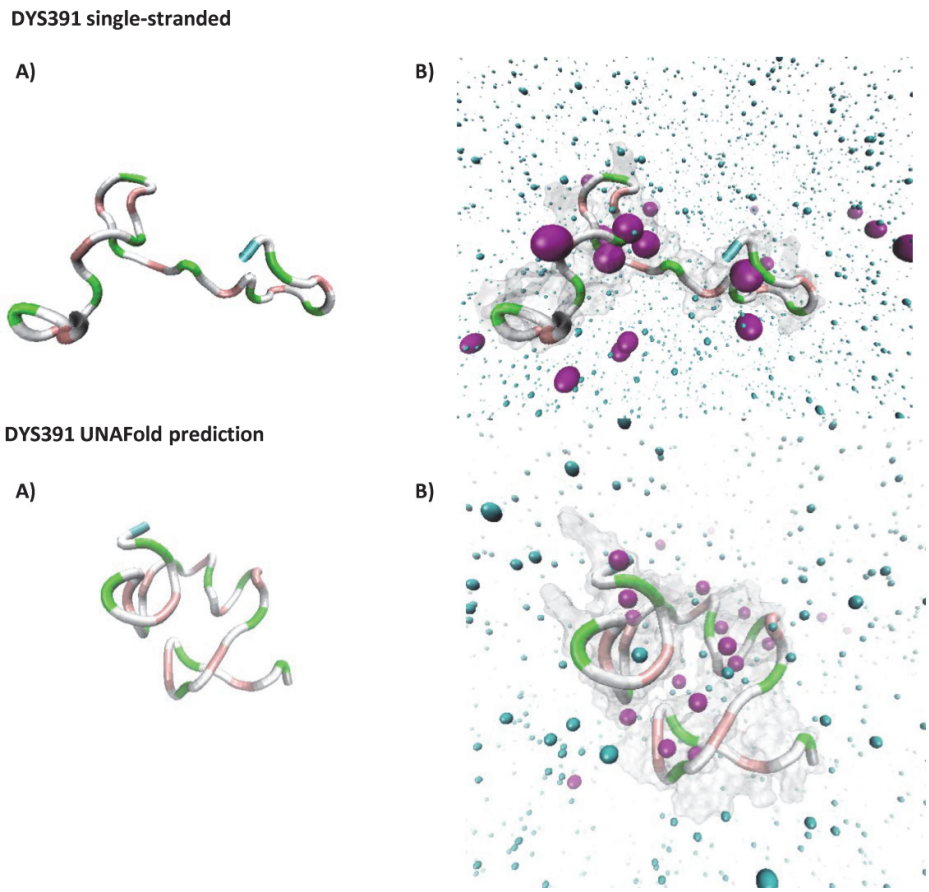


Figure 7: Snapshots of DYS391 12 ns molecular dynamics considering: A) Tube representation of full or partial folded hairpin(s) [DT-white, DC-green, DA-light pink]; B) Solvent (blue) and Na+ counterion (purple) distribution. DYS391 UNAFold prediction is associated with a supercoiled conformation.

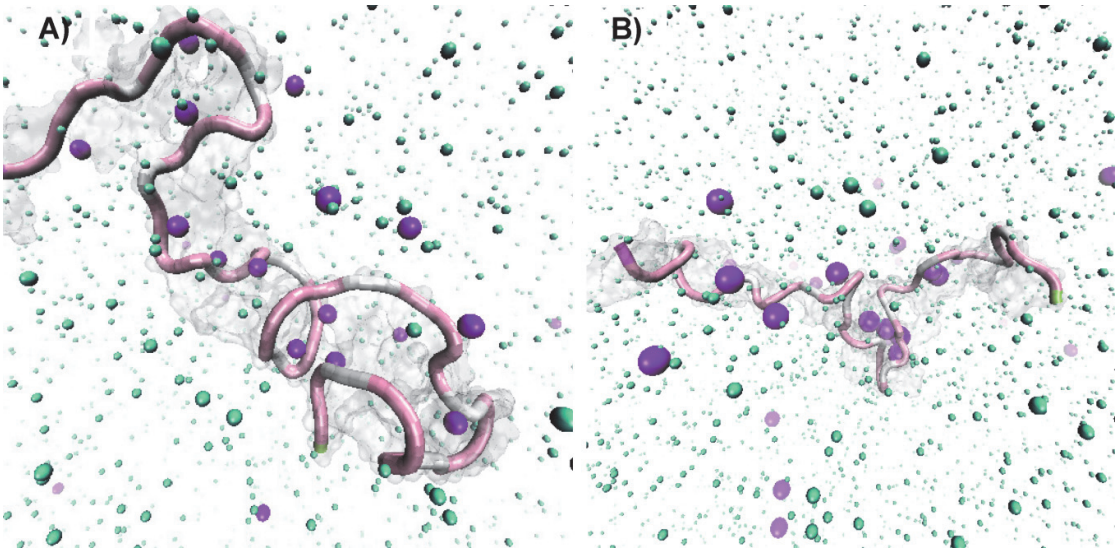


Figure 8: Snapshot of 12 ns molecular structures of: A) DYS531 single-stranded model tube representation of full or partial folded hairpin(s) (DT-white, DA-light pink); B) DYS531 UNAFold predicted model with median sized hairpin (DT-white, DA-light pink). The solvent (blue) and Na+ counterion (purple) distribution is represented in both models.

Supplementary Material

Table SI1: Table SI1: Y-STRs used for secondary structure prediction using UNAFold. Allele range, repeat motif, Genbank accession numbers and reference alleles of Y-STR locus. Repeat motif abreviations A,T,G,C,W,Y,R,S correspond respectively to adenine, thymine, guanine, cytosine, weak (A or T), pyrimidine, purine, strong (G or C) following the International Union of Pure and Applied Chemistry (IUPAC).

Marker Name	Allele Range* (repeat numbers)	Repeat Motif	GenBank Accession	Reference Allele	Hairpin prediction
DYS19A	10-19	$(TAGA)_3(TAGG)_1(TAGA)_{12}$	AC017019	15	Yes
DYS389 I	9-17	(TCTG) (TCTA) (TCTG) (TCTA)	AC004617	12	Yes
DYS389 II	24-34	(TCTG) (TCTA) (TCTG) (TCTA)	AC004617	29	Yes
DYS390	17-28	(TCTA) (TCTG)	AC011289	24	Yes
DYS391	6-14	TCTA	AC011302	11	Yes
DYS392	6-17	TAT	AC011745	13	Yes
DYS393	9-17	AGAT	AC006152	12	Yes
DYS388	10-18	ATT	AC004810	12	Yes
DYS425	10-14	TGT	AC095380	10	Yes
DYS426	10-12	GTT	AC007034	12	Yes
DYS434	9-12	TAAT (CTAT)	AC002992	10	Yes
DYS435	9-13	TGGA	AC002992	9	Yes
DYS436	9-15	GTT	AC005820	12	Yes
DYS442	10-14	$(TATC)_2(TGTC)_3(TATC)_{12}$	AC004810	17	Yes
DYS445	10-13	TTTA	AC009233	12	Yes
DYS447	22-29	TAA**W**A	AC005820	23	Yes
DYS448	20-26	AGAGAT	AC025227	22	Yes
DYS450	8-11	TTTTA	AC051663	9	Yes
DYS452	27-33	**Y**ATAC	AC010137	31	Yes
DYS454	10-12	AAAT	AC025731	11	Yes
DYS455	8-12	AAAT	AC012068	11	Yes
DYS456	13-18	AGAT	AC010106	15	Yes
DYS459 a/b	7-10	TAAA	AC010682	9	Yes
DYS460 (A7.1)	7-12	ATAG	AC009235	10	Yes
DYS462	8-14	TATG	AC007244	11	Yes
DYS485	10-18	TTA		16	Yes
DYS495	12-18	AAT	AC004474	15	Yes
DYS508	8-15	TATC	AC006462	11	Yes
DYS520	18-26	ATA**S**	AC007275	20	Yes
DYS522	8-17	GATA	AC007247	10	Yes
DYS531	9-13	AAAT		11	Yes
DYS533	9-14	ATCT	AC053516	12	Yes
DYS565	9-14	ATAA	AC010726	12	Yes
DYS573	8-11	TTTA		10	Yes
DYS594	9-14	AAATA	AC010137	10	Yes
DYS635 (C4)	17-27	T**S**TA compound	AC004772	23	Yes

Table SI2: Molecular dynamics models of Y-STRs: Single-stranded STR (SS); UNAFold predicted STR (UF); UNAFold predicted loop in different regions (UFR). The connected bases are represented by position number in STR like for example: "12b21b-13b20b" where there are base pairs between nucleotide 12 and 21, 13 and 20.

Y-STR	Sequence	Allele length	Models	Total number of models
DYS19A	$TAGA_3TAGG_1TAGA_{11}$	15	DYS19A-SS (DYS19A-SS_min) DYS19A-4b13b-5b12b-UF DYS19A-12b21b-13b20b-UFR DYS19A-20b29b-21b28b-UFR DYS19A-24b33b-25b32b-UFR DYS19A-28b37b-29b36b-UFR DYS19A-32b41b-33b40b-UFR DYS19A-36b45b-37b44b-UFR DYS19A-40b49b-41b48b-UFR DYS19A-44b53b-45b52b-UFR DYS19A-48b57b-49b56b-UFR	11
DYS391	$TCTA_{10}$	10	DYS391TCTA10-SS (DYS391TCTA10-SS_min) DYS391TCTA10-4b13b-5b12b-UF DYS391TCTA10-8b17b-9b16b-UFR DYS391TCTA10-12b21b-13b20b-UFR DYS391TCTA10-16b25b-17b24b-UFR DYS391TCTA10-20b29b-21b28b-UFR DYS391TCTA10-24b33b-25b32b-UFR DYS391TCTA10-28b37b-29b36b-UFR	8
DYS531	$AAAT_{11}$	11	DYS531AAAT11-SS (DYS531AAAT11-SS_min) DYS531AAAT11-3b12b-4b11b-UF DYS531AAAT11-7b16b-8b15b-UFR DYS531AAAT11-11b20b-12b19b-UFR DYS531AAAT11-15b24b-16b23b-UFR DYS531AAAT11-19b28b-20b27b-UFR DYS531AAAT11-23b32b-24b31b-UFR DYS531AAAT11-27b36b-28b35b-UFR DYS531AAAT11-31b40b-32b39b-UFR DYS531AAAT11-35b44b-36b43b-UFR	10

Figure SI1: Molecular dynamics simulation workflow.

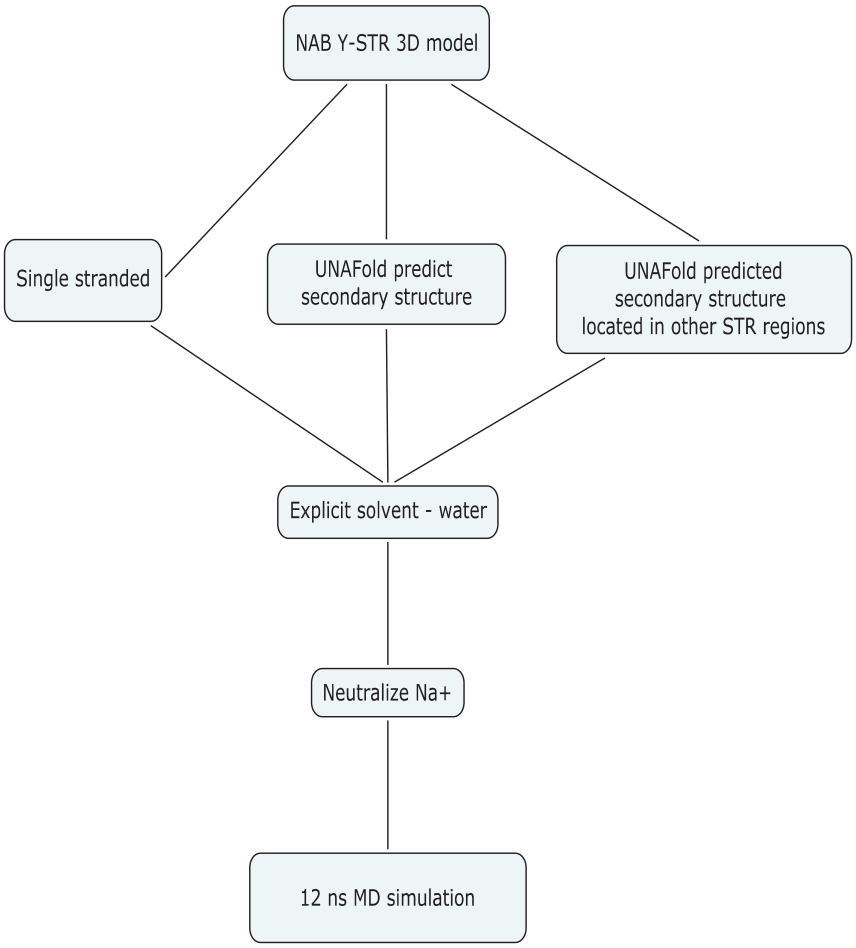
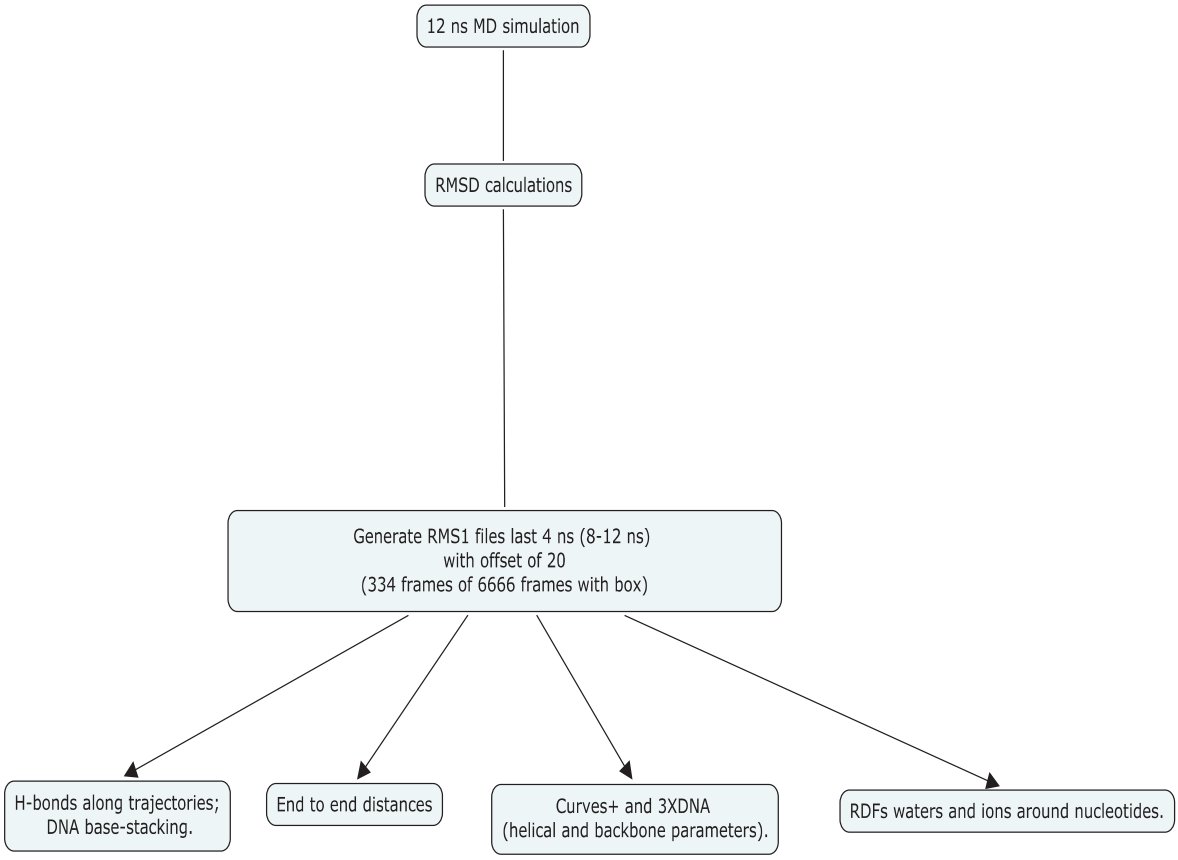


Figure SI2: Molecular Dynamics models analysis workflow.

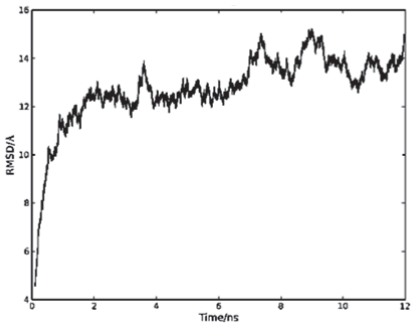


Root-Mean-Square deviation (RMSD) graphs

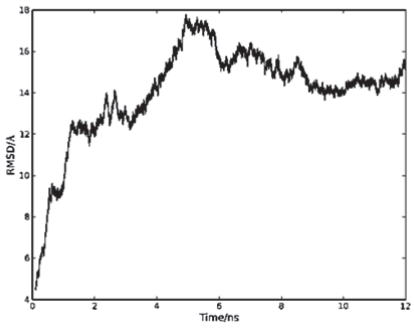
DYS19A (Imperfect STR, 60 nucleotides, 15 repeats):

Graphs SI1-SI11: RMSD graphs of backbone atoms for DYS19A MD simulation.

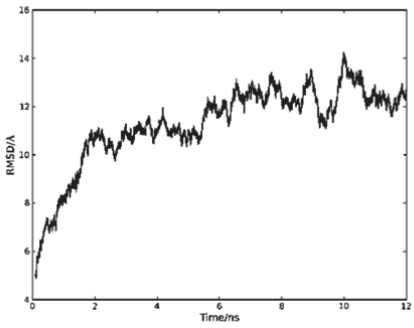
DYS19A-SS



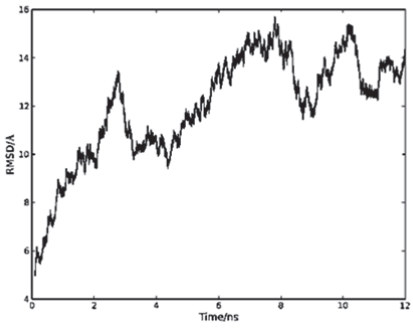
DYS19A-4b13b-5b12b-UF



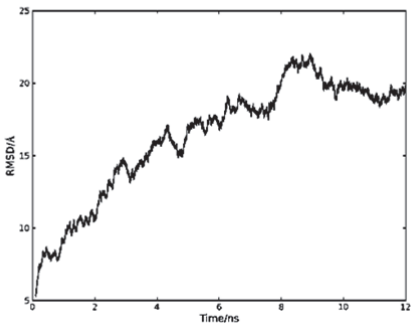
DYS19A-12b21b-13b20b-UFR



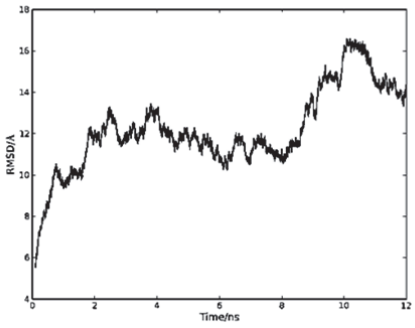
DYS19A-20b29b-21b28b-UFR



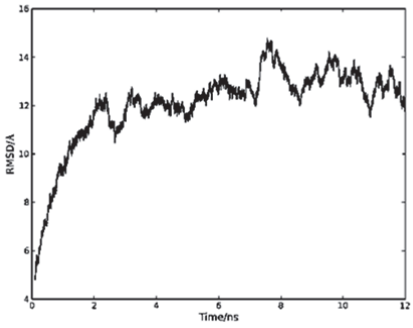
DYS19A-24b33b-25b32b-UFR



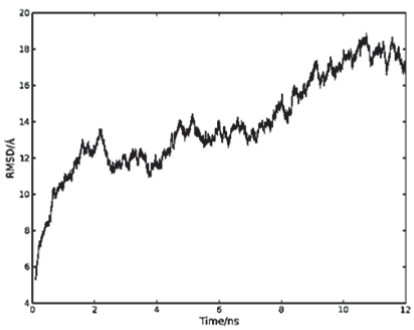
DYS19A-28b37b-29b36b-UFR



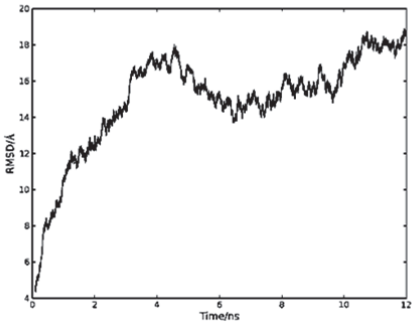
DYS19A-32b41b-33b40b-UFR



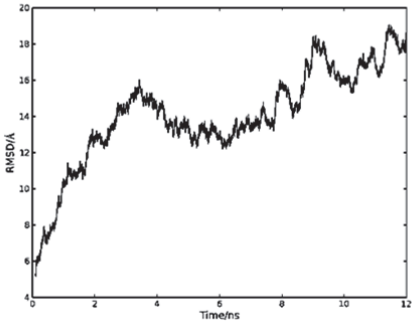
DYS19A-36b45b-37b44b-UFR



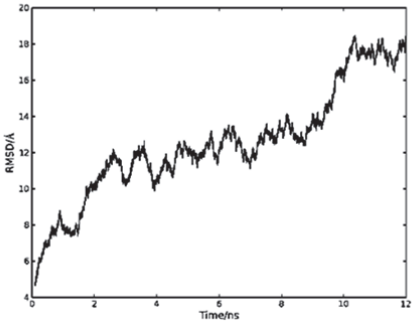
DYS19A-40b49b-41b48b-UFR



DYS19A-44b53b-45b52b-UFR



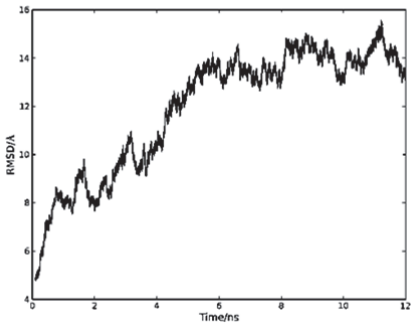
DYS19A-48b57b-49b56b-UFR



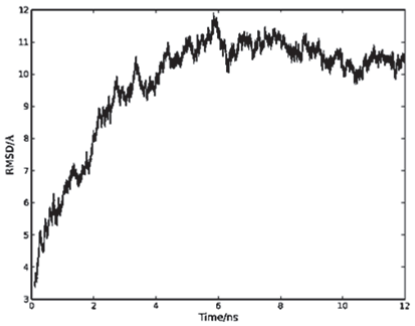
DYS391 (perfect STR, 40 nucleotides, 10 repeats):

Graphs SI12-SI19: RMSD graphs of backbone atoms for DYS391 MD simulation.

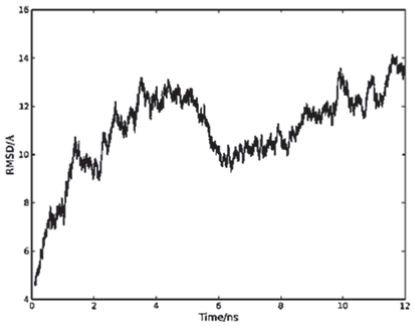
DYS391TCTA10-SS



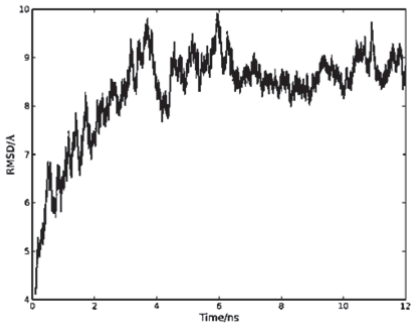
DYS391TCTA10-4b13b-5b12b-UF



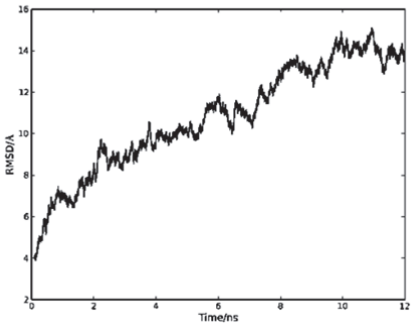
DYS391TCTA10-8b17b-9b16b-UFR



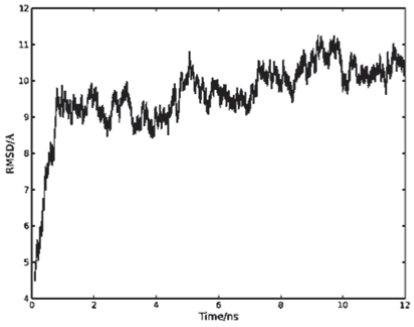
DYS391TCTA10-12b21b-13b20b-UFR



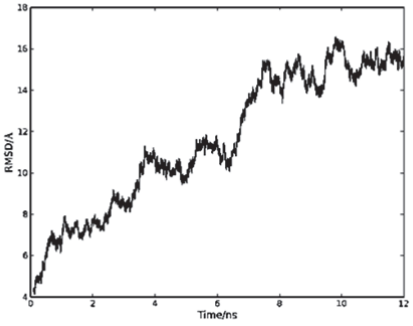
DYS391TCTA10-16b25b-17b24b-UFR



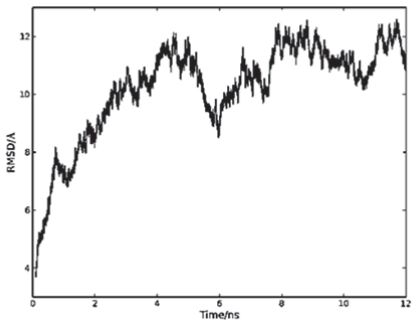
DYS391TCTA10-20b29b-21b28b-UFR



DYS391TCTA10-24b33b-25b32b-UFR



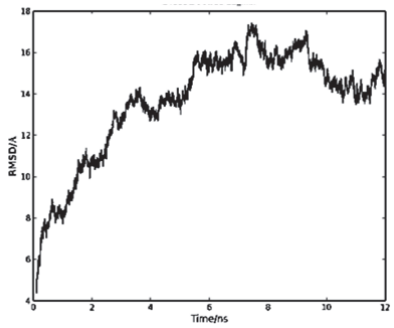
DYS391TCTA10-28b37b-29b36b-UFR



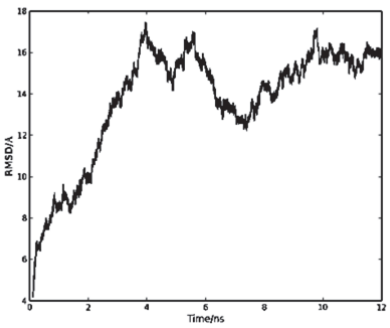
DYS531 (perfect STR, 44 nucleotides, 11 repeats):

Graphs SI20-SI29: RMSD graphs of backbone atoms for DYS531 MD simulation.

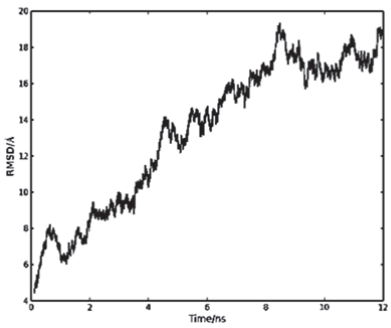
DYS531AAAT11-SS



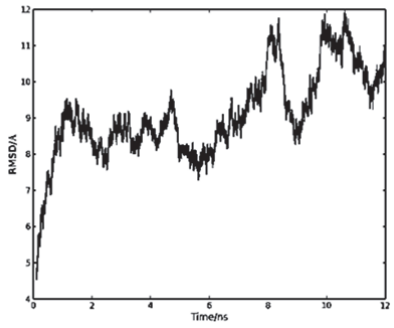
DYS531AAAT11-3b12b-4b11b-UF



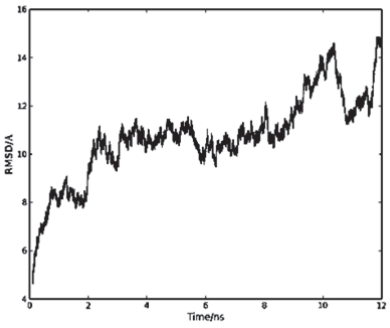
DYS531AAAT11-7b16b-8b15b-UFR



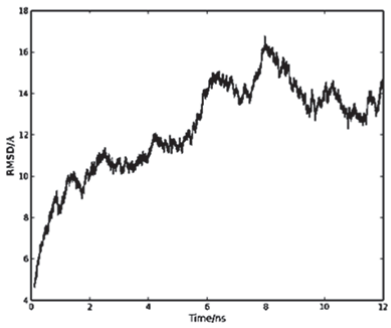
DYS531AAAT11-11b20b-12b19b-UFR



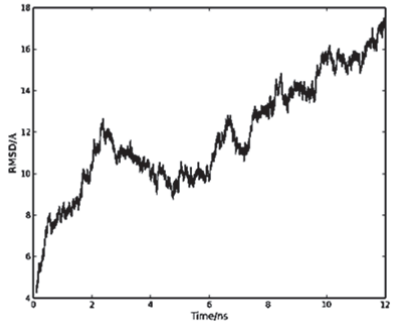
DYS531AAAT11-15b24b-16b23b-UFR



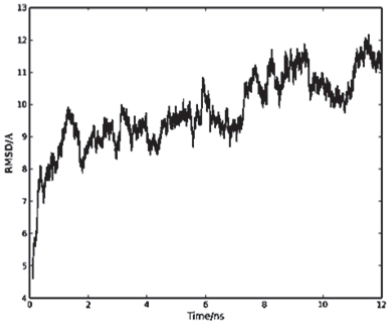
DYS531AAAT11-19b28b-20b27b-UFR



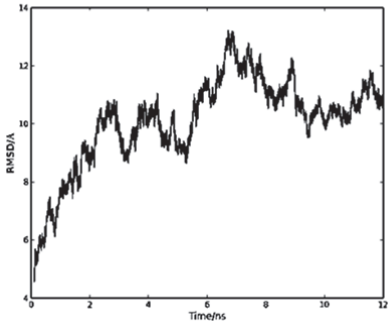
DYS531AAAT11-23b32b-24b31b-UFR



DYS531AAAT11-27b36b-28b35b-UFR



DYS531AAAT11-31b40b-32b39b-UFR



DYS531AAAT11-35b44b-36b43b-UFR

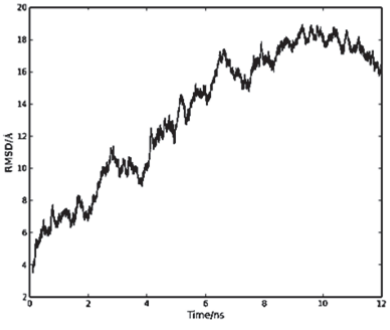


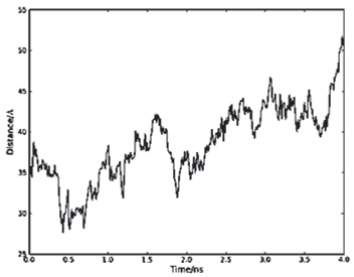
Table SI3: End to end distance basic statistics (first residue to last residue carbon atoms).

Y-STR model	Median/[Å)	Minimum/[Å)	Maximum/[Å)	Maximum - Minimum/[Å)	Standard deviation/[Å)
DYS19A-SS_min	38.53	27.61	51.73	24.12	4.64
DYS19A-4b13b-5b12b-UF	25.26	19.52	31.18	11.66	2.9
DYS19A-12b21b-13b20b-UFR	41.07	28.78	55.32	26.54	5.89
DYS19A-20b29b-21b28b-UFR	53.23	41.73	63.04	21.31	3.50
DYS19A-24b33b-25b32b-UFR	55.42	46.44	70.39	23.95	4.84
DYS19A-28b37b-29b36b-UFR	53.01	44.55	62.27	17.72	3.37
DYS19A-32b41b-33b40b-UFR	51.00	42.13	63.86	21.73	4.17
DYS19A-36b45b-37b44b-UFR	48.20	38.28	59.91	21.63	5.41
DYS19A-40b49b-41b48b-UFR	39.97	30.17	51.38	21.21	4.97
DYS19A-44b53b-45b52b-UFR	52.92	42.52	61.41	18.89	4.94
DYS19A-48b57b-49b56b-UFR	37.89	21.59	50.89	29.3	7.26
	45.14 (Average)	34.85 (Average)	56.49 (Average)	21.64 (Average)	
DYS391TCTA10-SS_min	35.35	27.76	45.37	17.61	3.93
DYS391TCTA10-4b13b-5b12b-UF	43.31	33.83	48.31	14.48	2.55
DYS391TCTA10-8b17b-9b16b-UFR	45.45	25.45	60.41	34.96	8.48
DYS391TCTA10-12b21b-13b20b-UFR	36.16	29.52	43.08	13.56	4.18
DYS391TCTA10-16b25b-17b24b-UFR	44.19	35.74	50.56	14.82	3.68
DYS391TCTA10-20b29b-21b28b-UFR	29.28	20.03	39.53	19.5	3.48
DYS391TCTA10-24b33b-25b32b-UFR	50.92	41.72	58.17	16.45	3.26
DYS391TCTA10-28b37b-29b36b-UFR	44.57	32.58	50.74	18.16	3.60
	41.15 (Average)	30.83 (Average)	49.52 (Average)	18.69 (Average)	
DYS531AAAT11-SS_min	51.01	42.59	58.35	15.76	3.54
DYS531AAAT11-3b12b-4b11b-UF	33.77	22.70	50.27	27.57	5.56
DYS531AAAT11-7b16b-8b15b-UFR	51.66	37.83	62.99	25.16	5.02
DYS531AAAT11-11b20b-12b19b-UFR	33.29	22.71	49.69	26.98	5.72
DYS531AAAT11-15b24b-16b23b-UFR	33.07	16.55	46.24	29.69	8.01
DYS531AAAT11-19b28b-20b27b-UFR	49.48	41.53	55.89	14.36	3.30
DYS531AAAT11-23b32b-24b31b-UFR	42.48	33.57	52.03	18.46	4.20
DYS531AAAT11-27b36b-28b35b-UFR	36.12	27.66	43.30	15.64	3.00
DYS531AAAT11-31b40b-32b39b-UFR	34.34	23.01	45.88	22.87	6.02
DYS531AAAT11-35b44b-36b43b-UFR	19.24	16.84	22.31	5.47	0.92
	38.446 (Average)	28.499 (Average)	48.695 (Average)	20.196 (Average)	

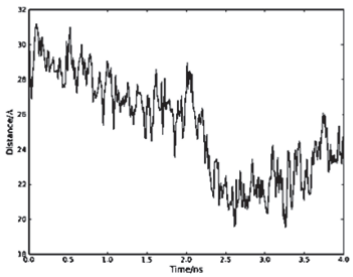
Distance from first residue carbon 1(C1' - 5' STR) to last residue carbon 1 (C1' - 3' STR) along the last 4 ns of trajectories.

Graphs SI30-SI40: End to end (C1' atoms) distance along trajectory (8-12 ns) for DYS19A.

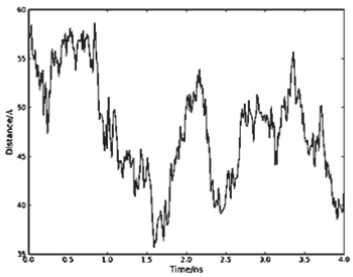
DYS19A-SS



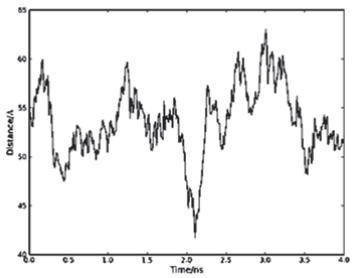
DYS19A-4b13b-5b12b-UF



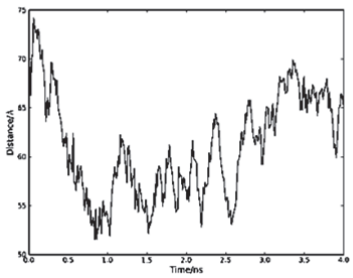
DYS19A-12b21b-13b20b-UFR



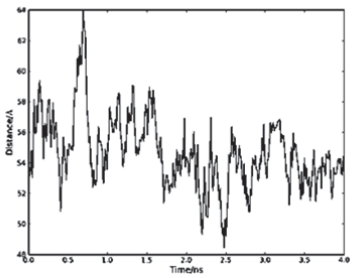
DYS19A-20b29b-21b28b-UFR



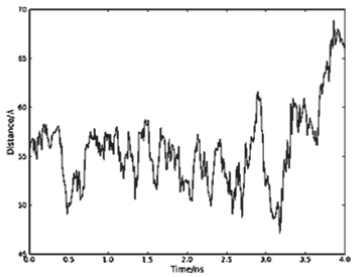
DYS19A-24b33b-25b32b-UFR



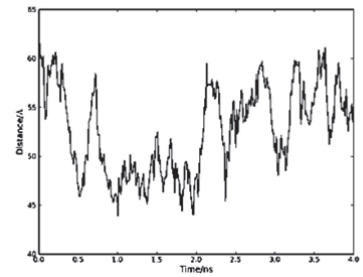
DYS19A-28b37b-29b36b-UFR



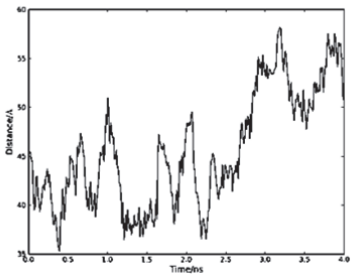
DYS19A-32b41b-33b40b-UFR



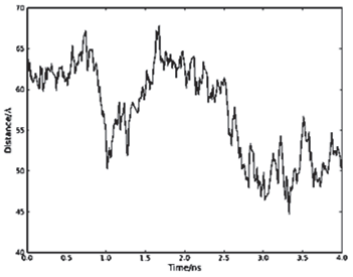
DYS19A-36b45b-37b44b-UFR



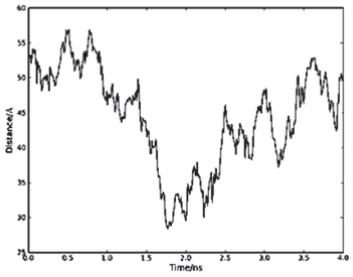
DYS19A-40b49b-41b48b-UFR



DYS19A-44b53b-45b52b-UFR

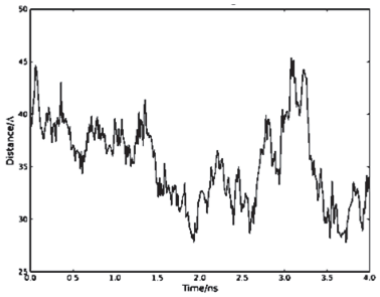


DYS19A-48b57b-49b56b-UFR

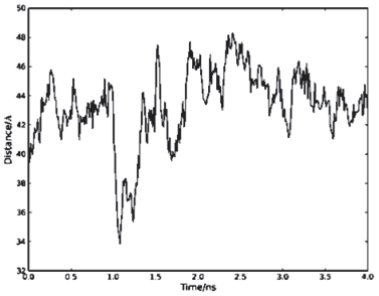


Graph SI41-SI48: End to end (C1' atoms) distance along trajectory (8-12 ns) for DYS391.

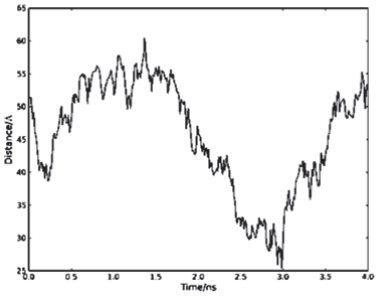
DYS391TCTA10-SS



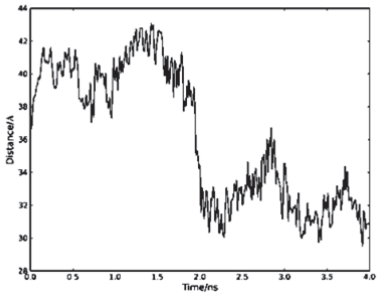
DYS391TCTA10-4b13b-5b12b-UF



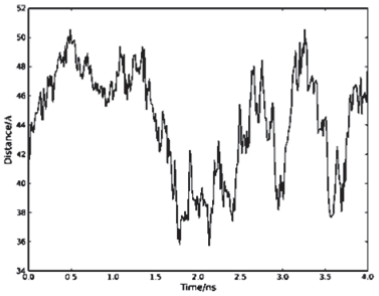
DYS391TCTA10-8b17b-9b16b-UFR



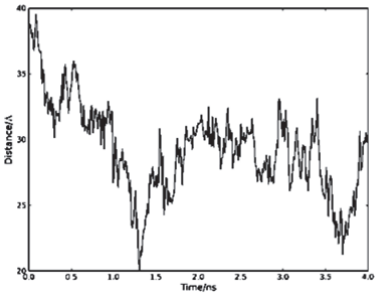
DYS391TCTA10-12b21b-13b20b-UFR



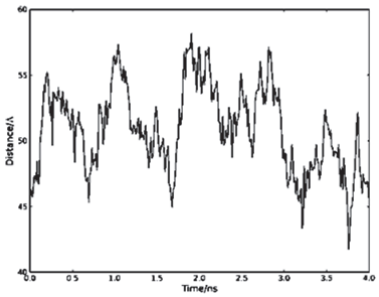
DYS391TCTA10-16b25b-17b24b-UFR



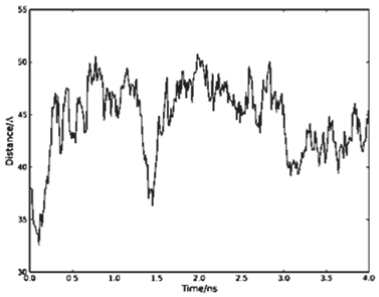
DYS391TCTA10-20b29b-21b28b-UFR



DYS391TCTA10-24b33b-25b32b-UFR

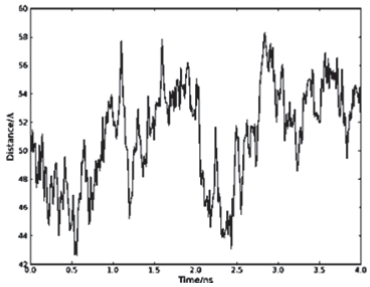


DYS391TCTA10-28b37b-29b36b-UFR

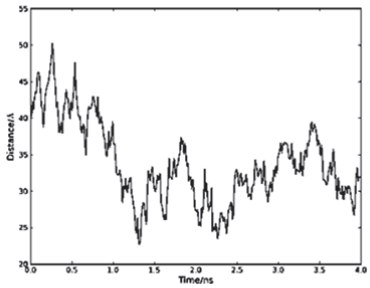


Graph SI49-SI58: End to end (C1' atoms) distance along trajectory (8-12 ns) for DYS531.

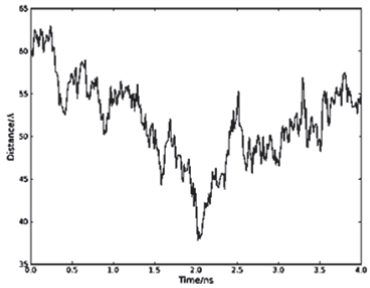
DYS531AAAT11-SS



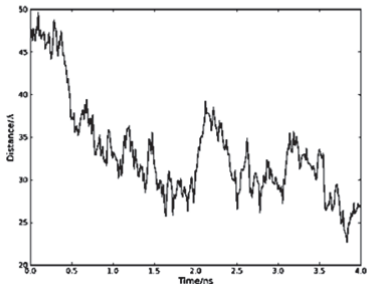
DYS531AAAT11-3b12b-4b11b-UF



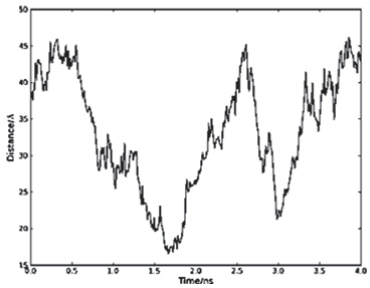
DYS531AAAT11-7b16b-8b15b-UFR



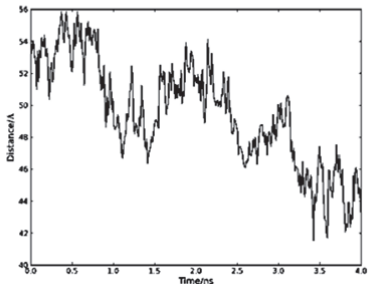
DYS531AAAT11-11b20b-12b19b-UFR



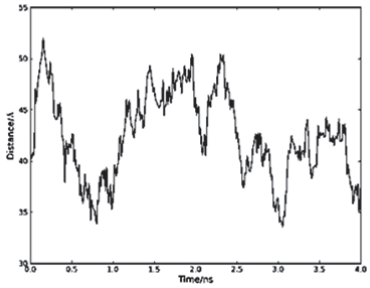
DYS531AAAT11-15b24b-16b23b-UFR



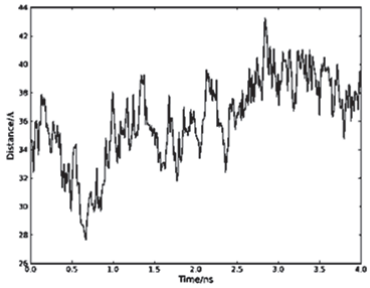
DYS531AAAT11-19b28b-20b27b-UFR



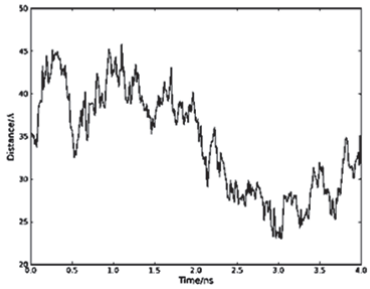
DYS531AAAT11-23b32b-24b31b-UFR



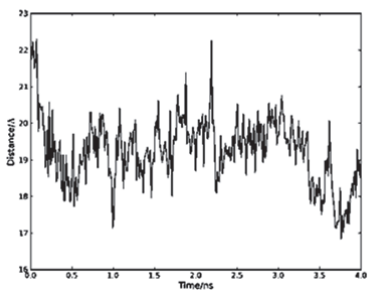
DYS531AAAT11-27b36b-28b35b-UFR



DYS531AAAT11-31b40b-32b39b-UFR



DYS531AAAT11-35b44b-36b43b-UFR



Tables SI4-SI14: Descriptive statistics of base pairing H-bonds in tested models of DYS19A Y-STR. Described values of H-bond percentage of occupancy during molecular dynamics (%occupied), distance (Å) of acceptor-donor H-bond (distance), angle of H-bond (angle), lifetime and maxocc as calculated by ptraj hbond function. Valid N corresponds to all H-bonds detected in the last 4 ns of the MD simulation.

	H-bonds DYS19A-SS_min				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	182	12.38	0.30	97.01	22.26
distance	182	4.24	2.92	4.98	0.57
angle	182	47.21	20.74	59.82	9.61
lifetime	182	11.56	4.00	190.70	24.73
maxocc	182	8.87	1.00	133.00	21.17

	H-bonds DYS19A-4b13b-5b12b-UF				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	239	8.89	0.30	99.70	18.02
distance	239	4.21	2.87	5.00	0.56
angle	239	47.41	16.97	59.99	9.08
lifetime	239	11.07	4.00	666.00	44.25
maxocc	239	8.16	1.00	239.00	24.32

	H-bonds DYS19A-12b21b-13b20b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	288	8.85	0.30	100.00	18.64
distance	288	4.19	2.87	4.98	0.59
angle	288	47.61	17.54	59.84	10.69
lifetime	288	25.22	4.00	1336.00	123.08
maxocc	288	12.23	1.00	334.00	40.98

	H-bonds DYS19A-20b29b-21b28b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	215	8.77	0.30	96.11	17.80
distance	215	4.21	2.90	4.99	0.56
angle	215	48.05	15.22	59.94	9.51
lifetime	215	10.51	4.00	287.00	28.78
maxocc	215	7.59	1.00	217.00	22.90

	H-bonds DYS19A-24b33b-25b32b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	274	8.25	0.30	100.00	17.40
distance	274	4.22	2.92	5.00	0.56
angle	274	46.32	6.10	59.84	9.93
lifetime	274	17.92	4.00	1336.00	114.18
maxocc	274	8.97	1.00	334.00	34.67

	H-bonds DYS19A-28b37b-29b36b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	249	6.95	0.30	100.00	15.06
distance	249	4.20	2.93	4.98	0.57
angle	249	47.05	14.89	59.72	9.66
lifetime	249	13.19	4.00	1336.00	84.87
maxocc	249	6.57	1.00	334.00	23.82

	H-bonds DYS19A-32b41b-33b40b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	245	7.74	0.30	97.60	15.22
distance	245	4.28	2.75	4.99	0.55
angle	245	46.38	20.16	59.84	10.22
lifetime	245	10.26	4.00	326.00	25.63
maxocc	245	8.07	1.00	253.00	25.76

	H-bonds DYS19A-36b45b-37b44b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	231	4.94	0.30	77.25	9.65
distance	231	4.20	2.87	4.98	0.52
angle	231	48.59	16.06	59.93	8.75
lifetime	231	6.95	4.00	86.00	8.37
maxocc	231	4.26	1.00	53.00	8.61

	H-bonds DYS19A-40b49b-41b48b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	203	5.84	0.30	92.51	13.57
distance	203	4.22	2.98	4.99	0.52
angle	203	48.18	20.28	59.97	8.65
lifetime	203	8.99	4.00	538.00	38.16
maxocc	203	4.48	1.00	137.00	14.13

	H-bonds DYS19A-44b53b-45b52b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	207	6.82	0.30	99.10	15.28
distance	207	4.31	2.96	4.99	0.51
angle	207	49.42	17.21	59.90	8.28
lifetime	207	8.75	4.00	331.00	26.55
maxocc	207	6.22	1.00	193.00	22.60

	H-bonds DYS19A-48b57b-49b56b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	208	8.21	0.30	99.40	16.66
distance	208	4.18	2.85	4.98	0.53
angle	208	47.69	13.89	59.95	9.92
lifetime	208	10.61	4.00	442.70	34.87
maxocc	208	7.04	1.00	144.00	20.87

Tables SI15-SI22: Descriptive statistics of base pairing H-bonds in tested models of DYS391 Y-STR. Described values of H-bond percentage of occupancy during molecular dynamics (%occupied), distance (Å) of acceptor-donor H-bond (distance), angle of H-bond (angle), lifetime and maxocc as calculated by ptraj hbond function. Valid N corresponds to all H-bonds detected in the last 4 ns of the MD simulation.

	H-bonds DYS391TCTA10-SS_min				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	147	5.20	0.30	77.54	10.30
distance	147	4.20	2.92	4.99	0.54
angle	147	45.81	12.81	59.94	10.56
lifetime	147	9.26	4.00	85.70	12.28
maxocc	147	6.22	1.00	114.00	14.41

	H-bonds DYS391TCTA10-4b13b-5b12b-UF				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	77	7.05	0.30	80.84	15.52
distance	77	4.27	2.94	4.97	0.50
angle	77	46.63	15.88	59.50	9.31
lifetime	77	6.96	4.00	45.70	7.19
maxocc	77	5.18	1.00	77.00	11.47

	H-bonds DYS391TCTA10-8b17b-9b16b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	123	7.77	0.30	97.31	15.36
distance	123	4.21	2.80	5.00	0.57
angle	123	46.81	18.47	59.73	9.93
lifetime	123	9.29	4.00	162.50	18.19
maxocc	123	8.44	1.00	194.00	24.90

	H-bonds DYS391TCTA10-12b21b-13b20b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	132	10.80	0.30	87.43	18.57
distance	132	4.19	3.00	5.00	0.56
angle	132	45.82	19.08	59.15	9.93
lifetime	132	11.66	4.00	149.30	19.99
maxocc	132	12.06	1.00	166.00	28.37

	H-bonds DYS391TCTA10-16b25b-17b24b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	144	7.12	0.30	97.31	15.79
distance	144	4.21	2.83	4.98	0.50
angle	144	47.60	14.35	59.79	9.09
lifetime	144	9.50	4.00	325.00	28.54
maxocc	144	5.80	1.00	166.00	16.94

	H-bonds DYS391TCTA10-20b29b-21b28b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	143	8.19	0.30	98.80	17.75
distance	143	4.19	2.96	4.97	0.53
angle	143	47.27	19.24	59.64	9.30
lifetime	143	10.81	4.00	330.00	30.50
maxocc	143	8.58	1.00	174.00	23.96

	H-bonds DYS391TCTA10-24b33b-25b32b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	89	8.90	0.30	97.90	19.03
distance	89	4.16	2.88	4.98	0.53
angle	89	47.78	22.79	59.68	8.60
lifetime	89	8.02	4.00	163.50	17.71
maxocc	89	6.06	1.00	188.00	20.94

	H-bonds DYS391TCTA10-28b37b-29b36b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	144	13.35	0.30	100.00	21.5
distance	144	4.27	2.89	4.99	0.6
angle	144	45.68	10.49	59.94	10.6
lifetime	144	29.17	4.00	1336.00	157.6
maxocc	144	14.56	1.00	334.00	46.5

Tables SI23-SI32: Descriptive statistics of base pairing H-bonds in tested models of DYS531 Y-STR. Described values of H-bond percentage of occupancy during molecular dynamics (%occupied), distance (Å) of acceptor-donor H-bond (distance), angle of H-bond (angle), lifetime and maxocc as calculated by ptraj hbond function. Valid N corresponds to all H-bonds detected in the last 4 ns of the MD simulation.

	H-bonds DYS531AAAT11-SS_min				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	177	11.38	0.30	100.00	21.86
distance	177	4.32	2.77	4.99	0.55
angle	177	46.45	16.67	59.85	10.42
lifetime	177	19.76	4.00	1336.00	105.27
maxocc	177	12.63	1.00	334.00	41.68

	H-bonds DYS531AAAT11-3b12b-4b11b-UF				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	152	5.35	0.30	91.92	12.32
distance	152	4.22	2.87	4.99	0.51
angle	152	48.12	7.60	59.62	9.71
lifetime	152	6.74	4.00	136.40	11.42
maxocc	152	4.04	1.00	133.00	12.06

	H-bonds DYS531AAAT11-7b16b-8b15b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	162	8.66	0.30	84.43	14.66
distance	162	4.24	2.99	4.97	0.52
angle	162	45.98	9.76	60.00	10.55
lifetime	162	10.08	4.00	202.00	20.48
maxocc	162	7.32	1.00	110.00	16.68

	H-bonds DYS531AAAT11-11b20b-12b19b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	168	5.39	0.30	97.31	12.0
distance	168	4.22	3.11	4.99	0.5
angle	168	47.07	20.17	59.99	9.1
lifetime	168	7.34	4.00	144.40	11.7
maxocc	168	4.07	1.00	91.00	9.0

	H-bonds DYS531AAAT11-15b24b-16b23b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	113	8.15	0.30	90.72	16.08
distance	113	4.21	2.83	4.99	0.53
angle	113	49.61	18.45	59.95	8.79
lifetime	113	6.54	4.00	88.30	9.21
maxocc	113	3.69	1.00	54.00	7.53

	H-bonds DYS531AAAT11-19b28b-20b27b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	175	8.81	0.30	99.40	18.30
distance	175	4.25	2.80	4.98	0.53
angle	175	46.93	17.59	59.66	9.22
lifetime	175	13.88	4.00	442.70	51.72
maxocc	175	8.82	1.00	304.00	31.56

	H-bonds DYS531AAAT11-23b32b-24b31b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	177	7.54	0.30	100.00	13.02
distance	177	4.17	2.98	5.00	0.53
angle	177	47.02	17.89	59.60	9.86
lifetime	177	18.14	4.00	1336.00	105.85
maxocc	177	7.92	1.00	334.00	28.43

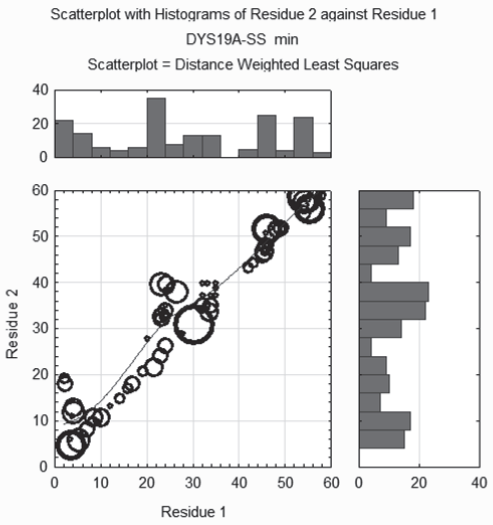
	H-bonds DYS531AAAT11-27b36b-28b35b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	157	9.40	0.30	99.10	18.52
distance	157	4.30	2.97	5.00	0.55
angle	157	48.39	17.06	59.97	8.77
lifetime	157	12.37	4.00	331.00	39.15
maxocc	157	7.86	1.00	173.00	23.42

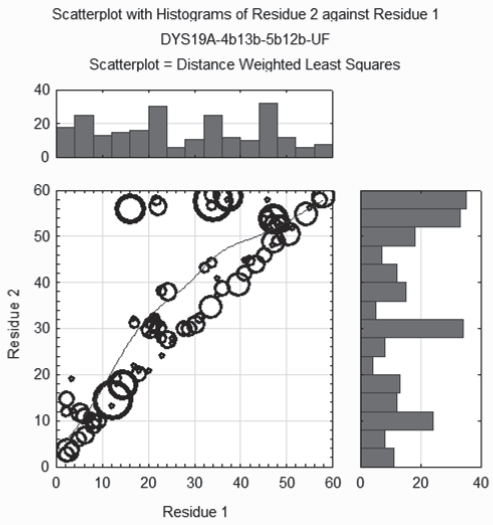
	H-bonds DYS531AAAT11-31b40b-32b39b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	175	7.02	0.30	98.80	15.42
distance	175	4.29	2.88	5.00	0.55
angle	175	46.46	19.98	59.71	9.11
lifetime	175	10.34	4.00	264.00	25.93
maxocc	175	6.84	1.00	194.00	21.76

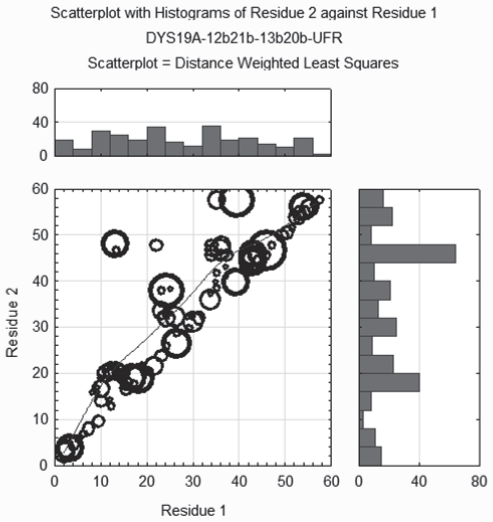
	H-bonds DYS531AAAT11-35b44b-36b43b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	184	6.25	0.30	70.06	10.24
distance	184	4.26	2.96	4.97	0.49
angle	184	47.42	17.83	59.95	9.01
lifetime	184	10.24	4.00	292.00	30.00
maxocc	184	5.78	1.00	82.00	12.08

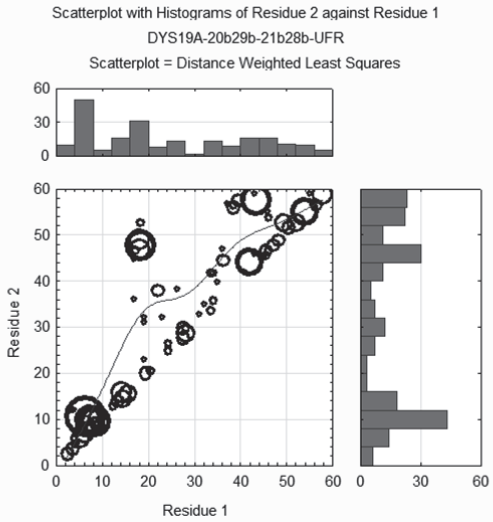
Scatterplots of nucleotide against nucleotide (position number) - base pairing H-bonds

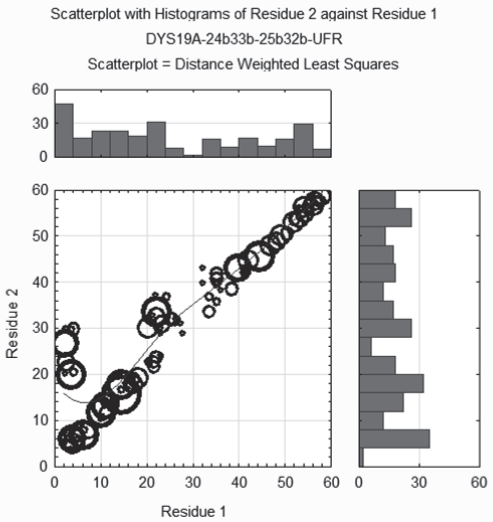
Graphs SI59-SI69: Scatterplots of nucleotide against nucleotide (position number) - base pairing H-bonds of DYS19A; Histograms categories represent the number of observations for motifs of four nucleotides.

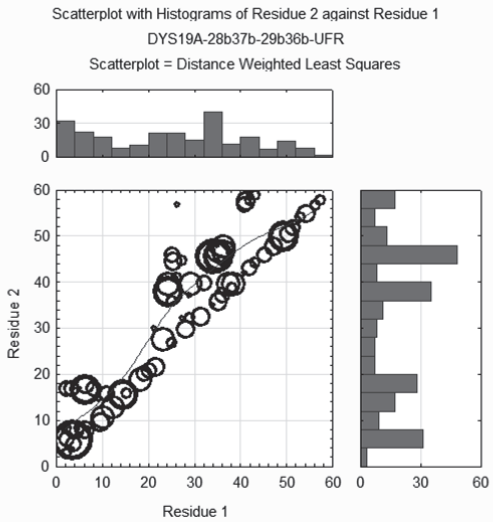

Scatterplot with Histograms of Residue 2 against Residue 1
DYS19A-SS min
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

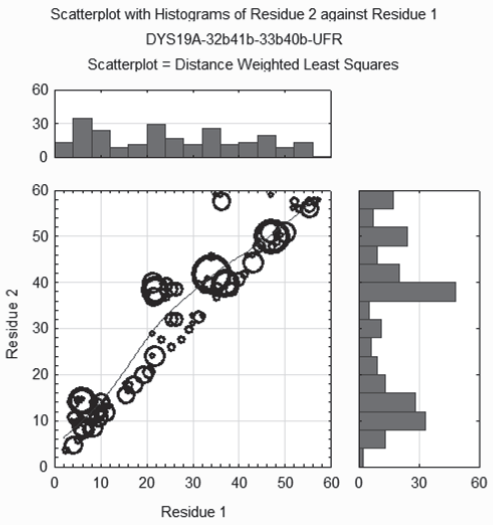

Scatterplot with Histograms of Residue 2 against Residue 1
DYS19A-4b13b-5b12b-UF
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

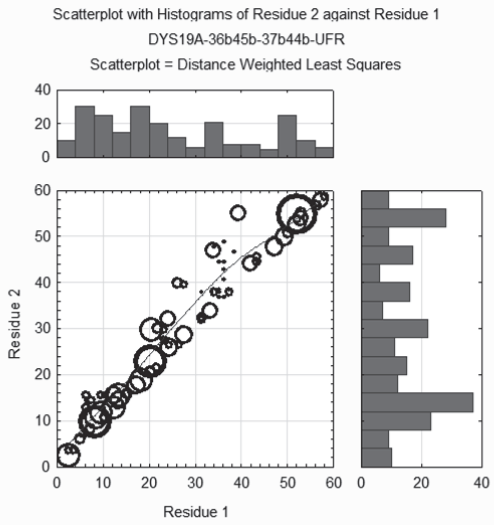

Scatterplot with Histograms of Residue 2 against Residue 1
DYS19A-12b21b-13b20b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

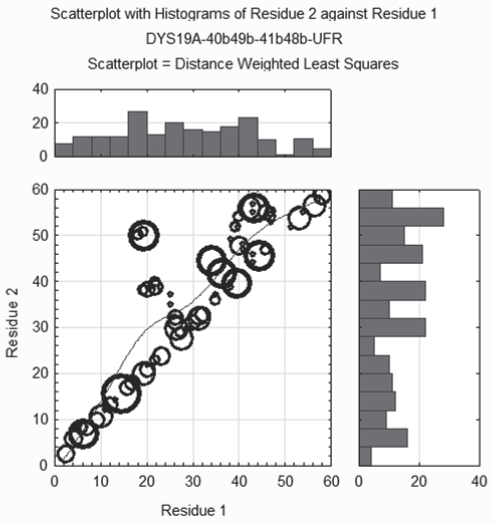

Scatterplot with Histograms of Residue 2 against Residue 1
DYS19A-20b29b-21b28b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

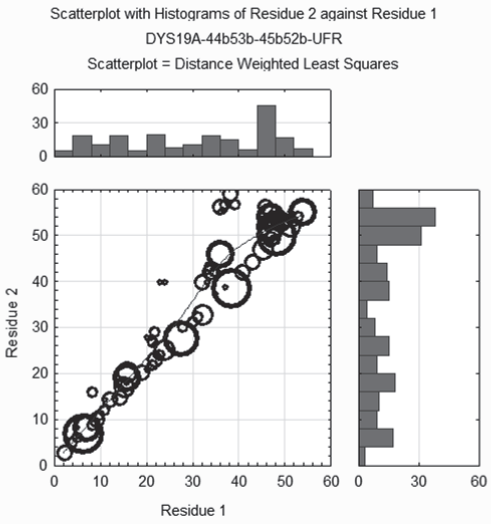

Scatterplot with Histograms of Residue 2 against Residue 1
DYS19A-24b33b-25b32b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

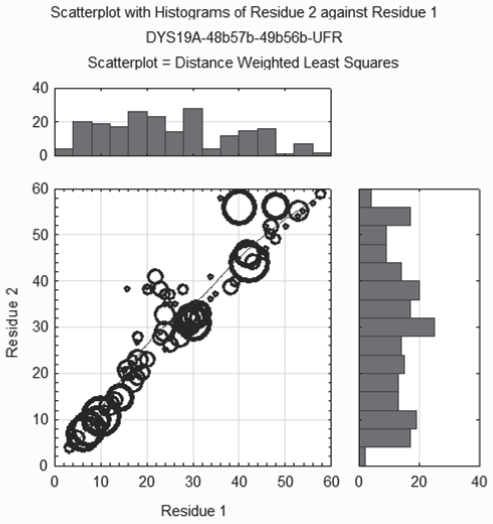

Scatterplot with Histograms of Residue 2 against Residue 1
DYS19A-28b37b-29b36b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1


Scatterplot with Histograms of Residue 2 against Residue 1
DYS19A-32b41b-33b40b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

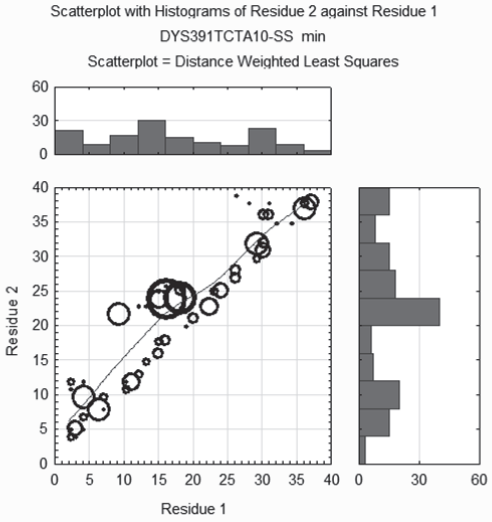

Scatterplot with Histograms of Residue 2 against Residue 1
DYS19A-36b45b-37b44b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

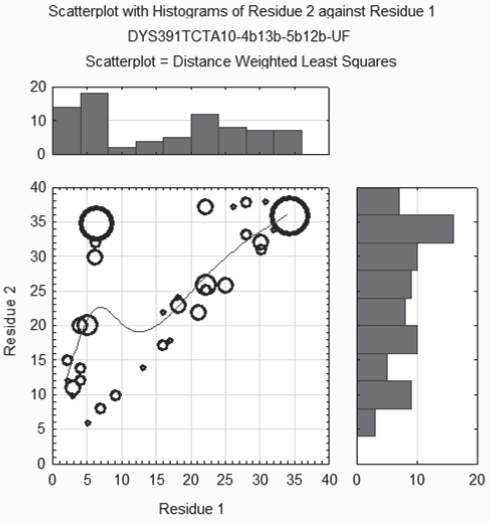

Scatterplot with Histograms of Residue 2 against Residue 1
DYS19A-40b49b-41b48b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

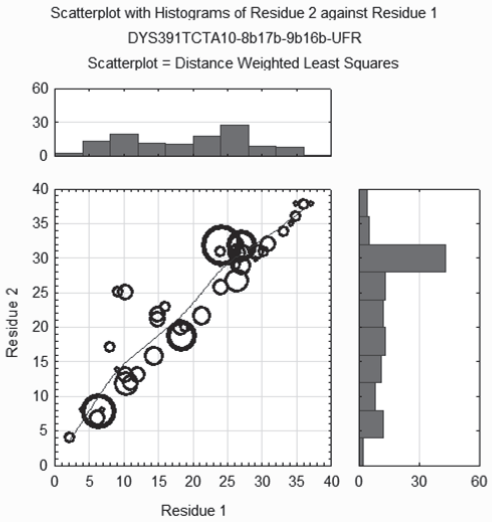

Scatterplot with Histograms of Residue 2 against Residue 1
DYS19A-44b53b-45b52b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

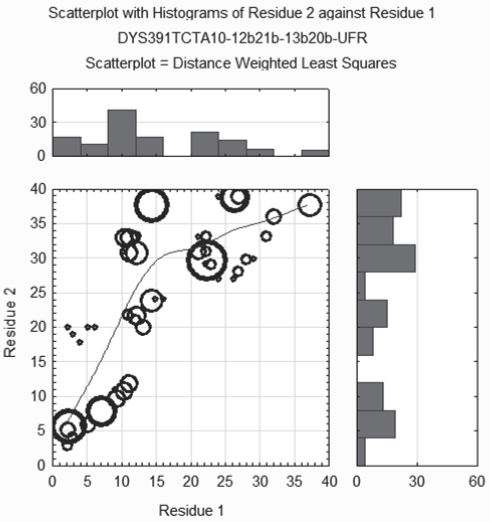

Scatterplot with Histograms of Residue 2 against Residue 1
DYS19A-48b57b-49b56b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

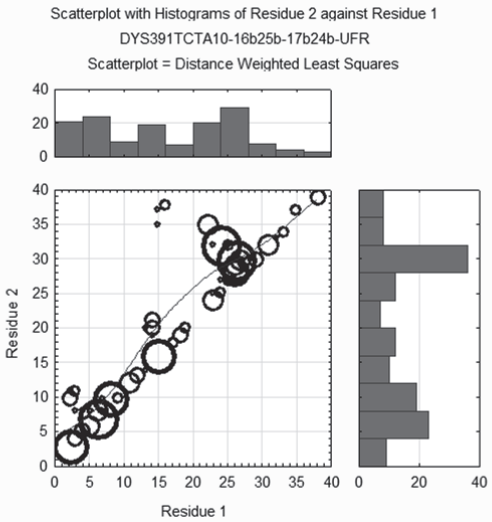
Graphs SI70-SI77: Scatterplots of nucleotide against nucleotide (position number) - base pairing of H-bonds of DYS391; Histograms categories represent the number of observations for motifs of four nucleotides.

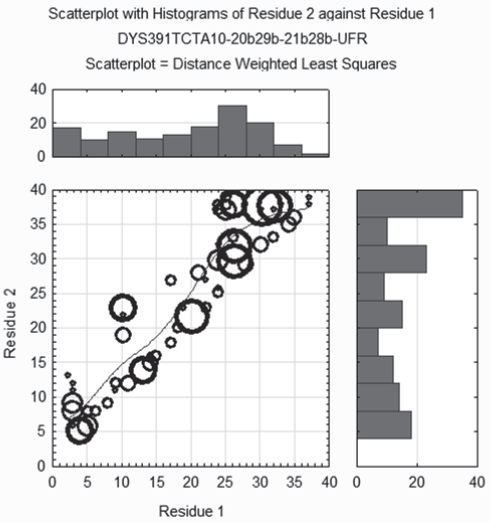

Scatterplot with Histograms of Residue 2 against Residue 1
DYS391TCTA10-SS min
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

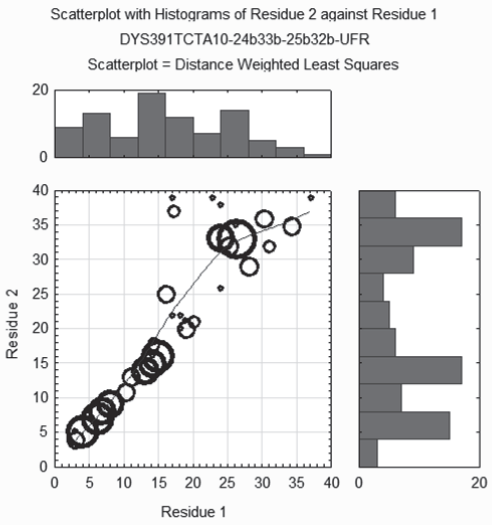

Scatterplot with Histograms of Residue 2 against Residue 1
DYS391TCTA10-4b13b-5b12b-UF
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

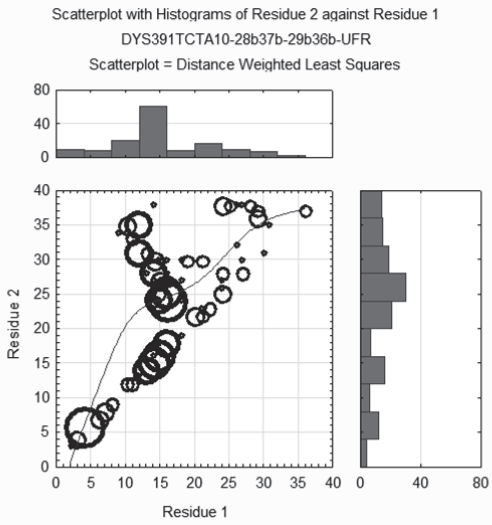

Scatterplot with Histograms of Residue 2 against Residue 1
DYS391TCTA10-8b17b-9b16b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1


Scatterplot with Histograms of Residue 2 against Residue 1
DYS391TCTA10-12b21b-13b20b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

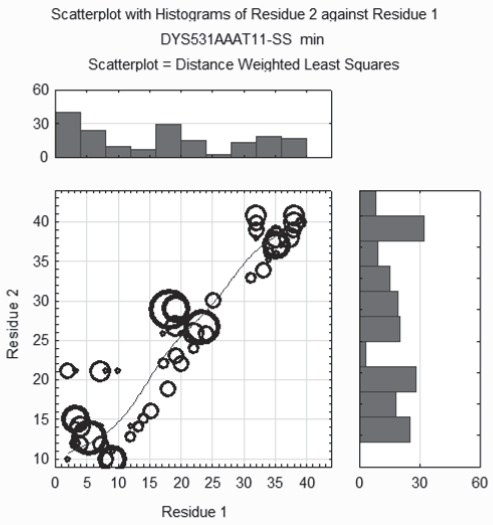

Scatterplot with Histograms of Residue 2 against Residue 1
DYS391TCTA10-16b25b-17b24b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

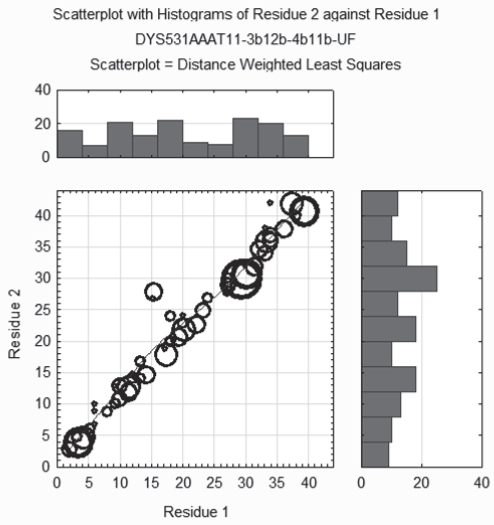

Scatterplot with Histograms of Residue 2 against Residue 1
DYS391TCTA10-20b29b-21b28b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

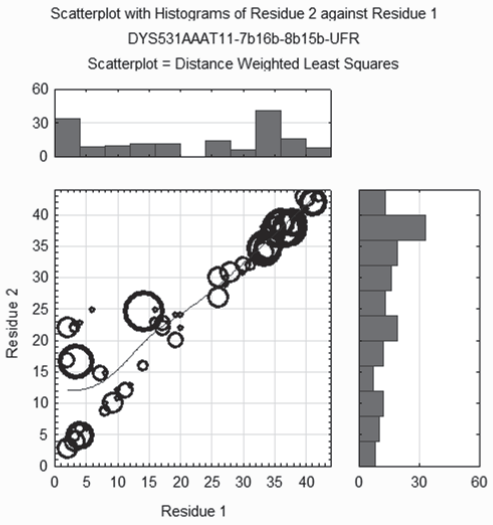

Scatterplot with Histograms of Residue 2 against Residue 1
DYS391TCTA10-24b33b-25b32b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1
0 5 10 15 20 25 30 35 40
0 20

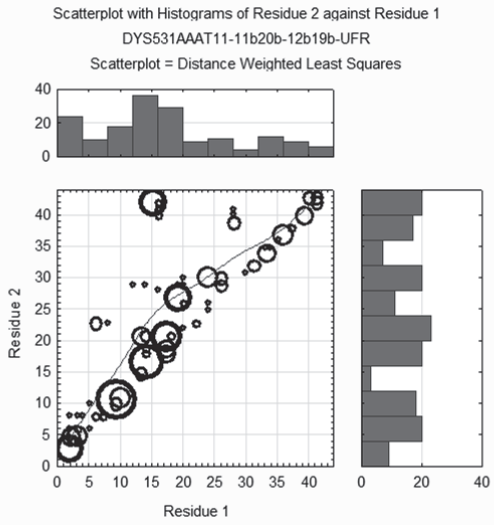

Scatterplot with Histograms of Residue 2 against Residue 1
DYS391TCTA10-28b37b-29b36b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1
0 5 10 15 20 25 30 35 40
0 40 80

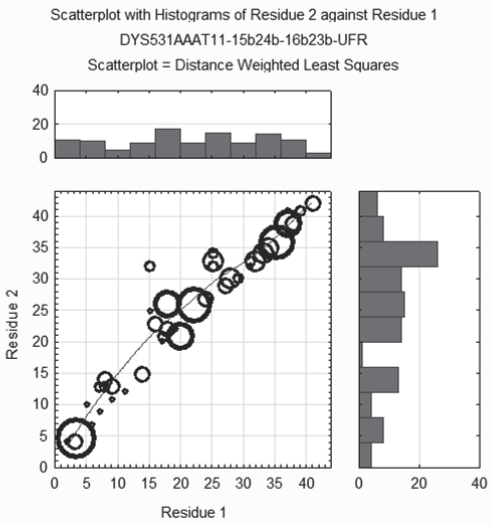
Graphs SI78-SI87: Scatterplots of nucleotide against nucleotide (position number) - base pairing H-bonds of DYS531; Histograms categories represent the number of observations for motifs of four nucleotides.

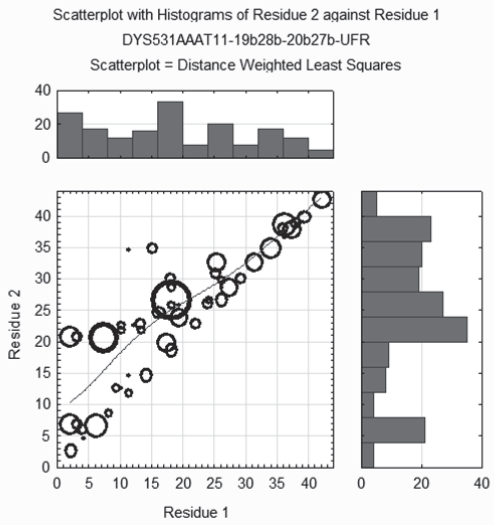

Scatterplot with Histograms of Residue 2 against Residue 1
DYS531AAAT11-SS min
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

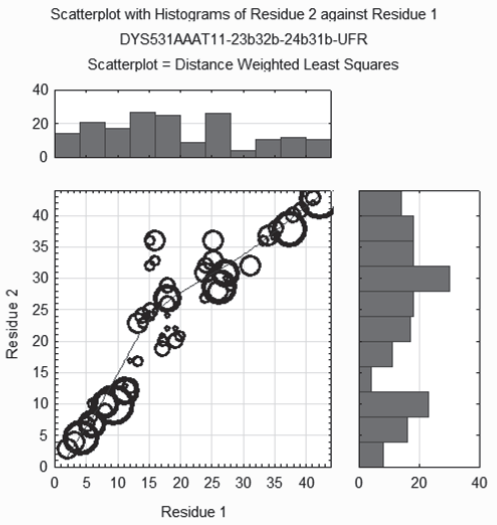

Scatterplot with Histograms of Residue 2 against Residue 1
DYS531AAAT11-3b12b-4b11b-UF
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

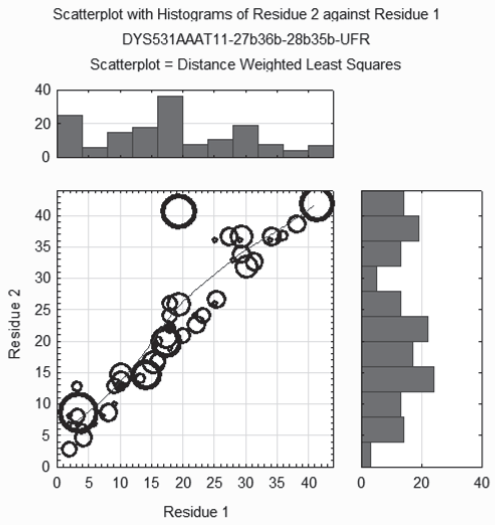

Scatterplot with Histograms of Residue 2 against Residue 1
DYS531AAAT11-7b16b-8b15b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

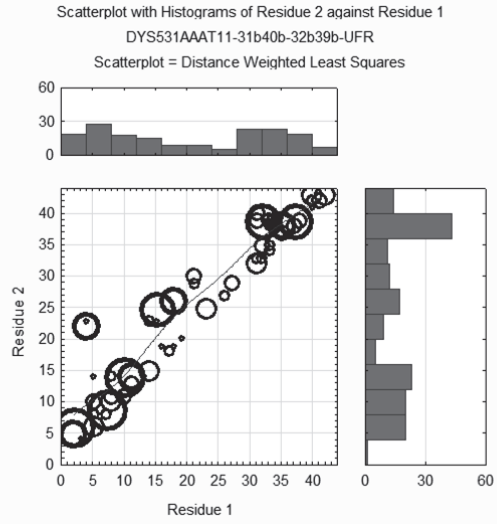

Scatterplot with Histograms of Residue 2 against Residue 1
DYS531AAAT11-11b20b-12b19b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

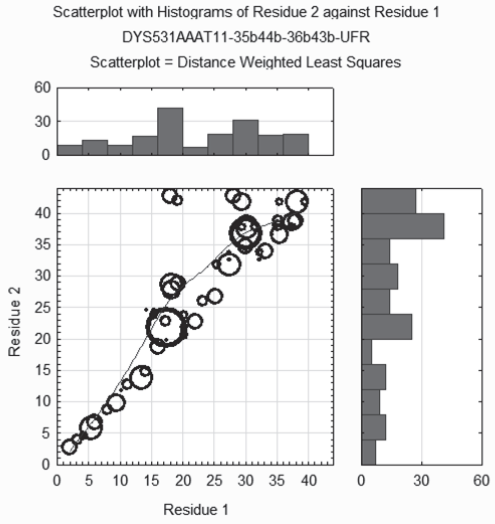

Scatterplot with Histograms of Residue 2 against Residue 1
DYS531AAAT11-15b24b-16b23b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1


Scatterplot with Histograms of Residue 2 against Residue 1
DYS531AAAT11-19b28b-20b27b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

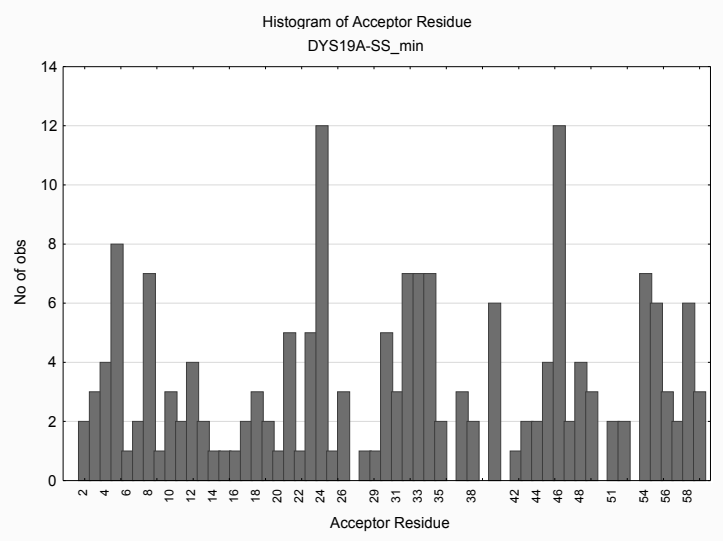

Scatterplot with Histograms of Residue 2 against Residue 1
DYS531AAAT11-23b32b-24b31b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

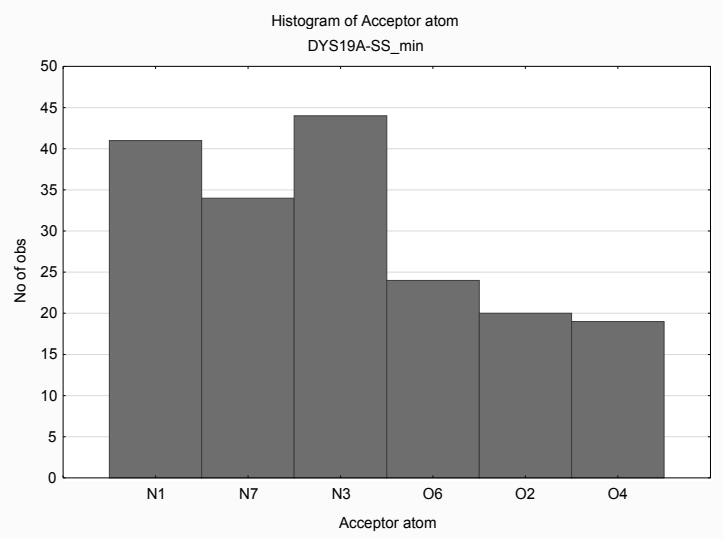

Scatterplot with Histograms of Residue 2 against Residue 1
DYS531AAAT11-27b36b-28b35b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

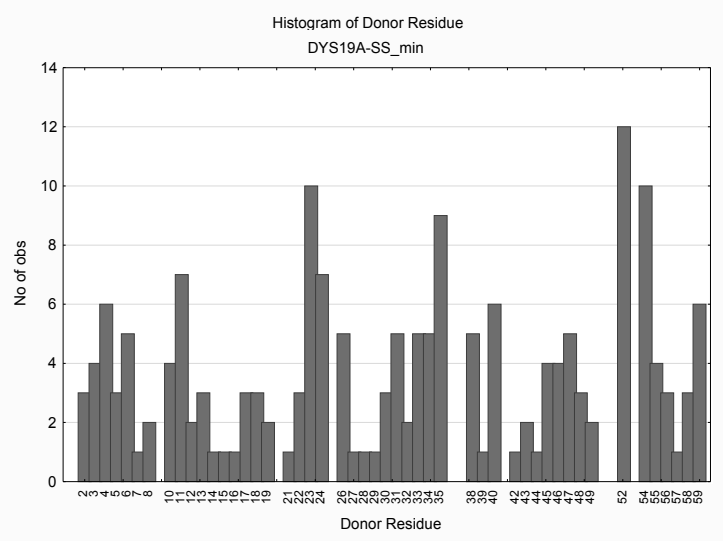

Scatterplot with Histograms of Residue 2 against Residue 1
DYS531AAAT11-31b40b-32b39b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

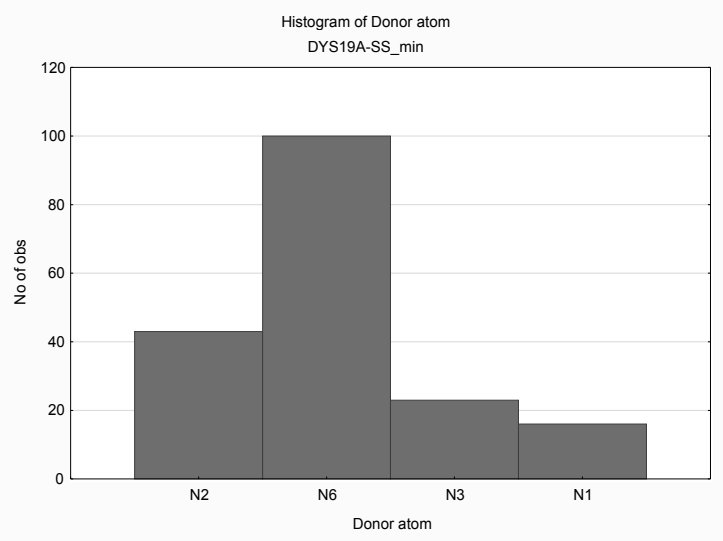

Scatterplot with Histograms of Residue 2 against Residue 1
DYS531AAAT11-35b44b-36b43b-UFR
Scatterplot = Distance Weighted Least Squares
Residue 2
Residue 1

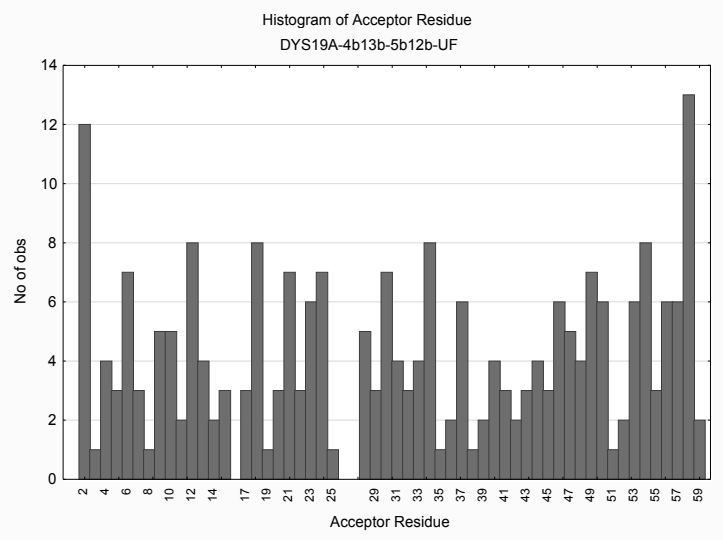
Graphs SI88-SI95: Histograms for hydrogen bonding residues and atoms (acceptors and donors) of DYS19A single-stranded and UNAFold predicted structure. Histogram of atoms considers the total number of H-bonds of the macromolecular model for each atom type. Histogram of residues considers the number of H-bonds of the macromolecular model for each residue.

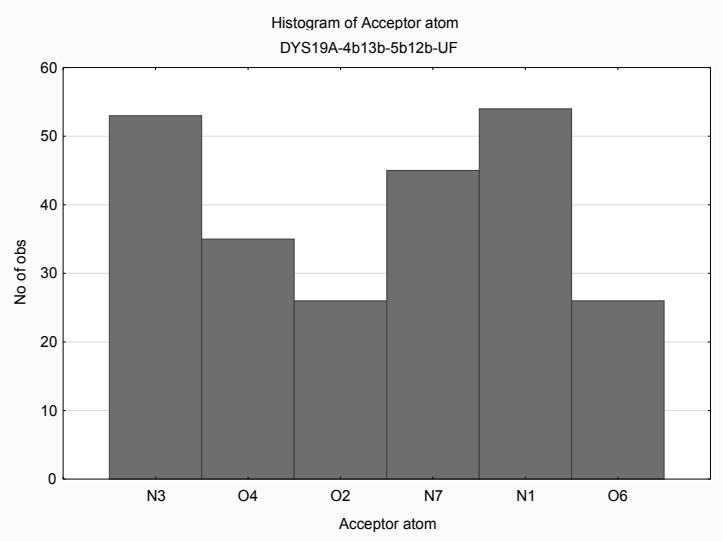


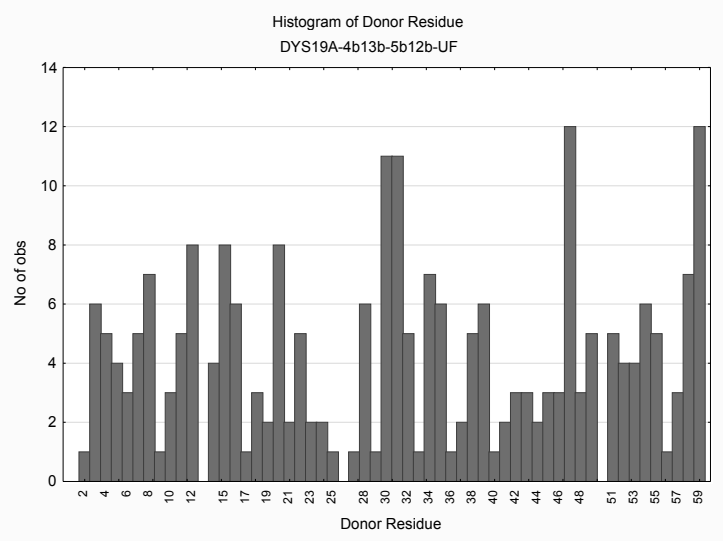


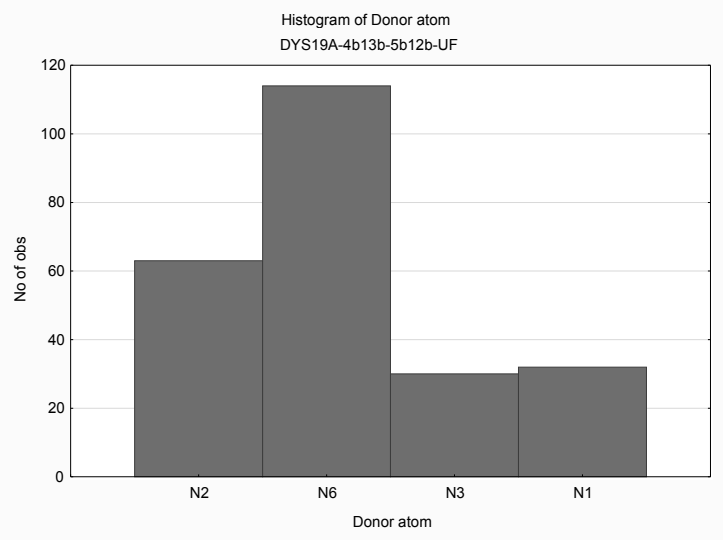

Histogram of Donor Residue
DYS19A-SS_min
No of obs
Donor Residue


Histogram of Donor atom
DYS19A-SS_min
No of obs
N2
N6
N3
N1
Donor atom

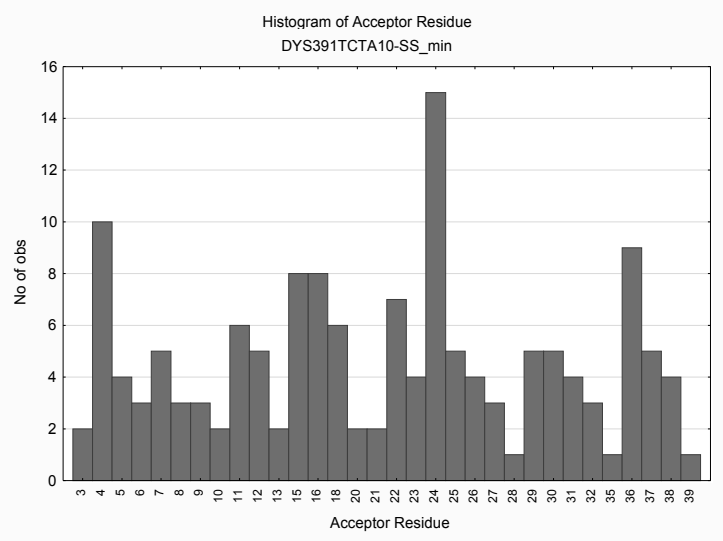

Histogram of Acceptor Residue
DYS19A-4b13b-5b12b-UF
No of obs
Acceptor Residue

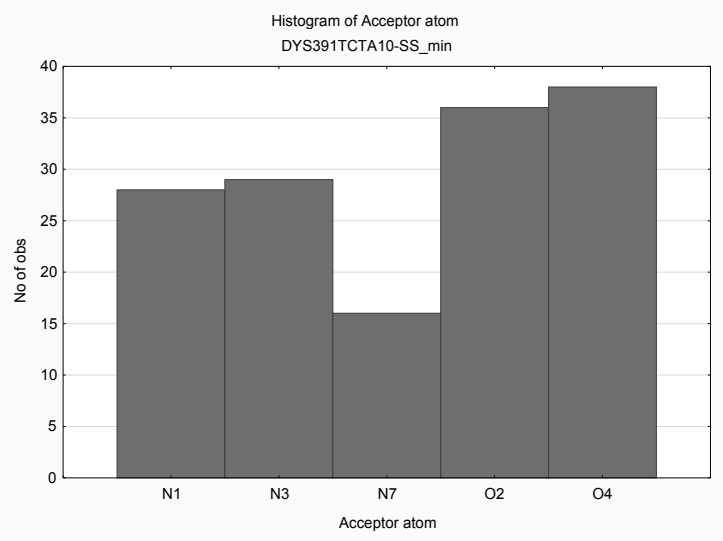

Histogram of Acceptor atom
DYS19A-4b13b-5b12b-UF
No of obs
0
10
20
30
40
50
60
N3
O4
O2
N7
N1
O6
Acceptor atom

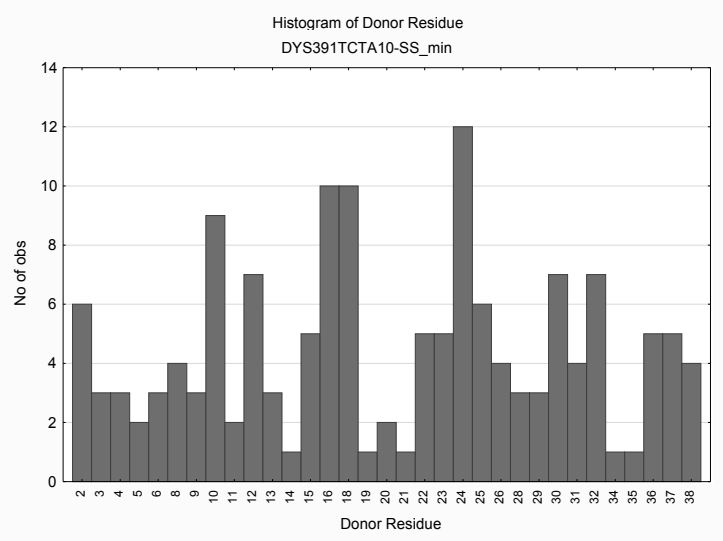

Histogram of Donor Residue
DYS19A-4b13b-5b12b-UF
No of obs
0
2
4
6
8
10
12
14
2
4
6
8
10
12
15
17
19
21
23
25
28
30
32
34
36
38
40
42
44
46
48
51
53
55
57
59
Donor Residue

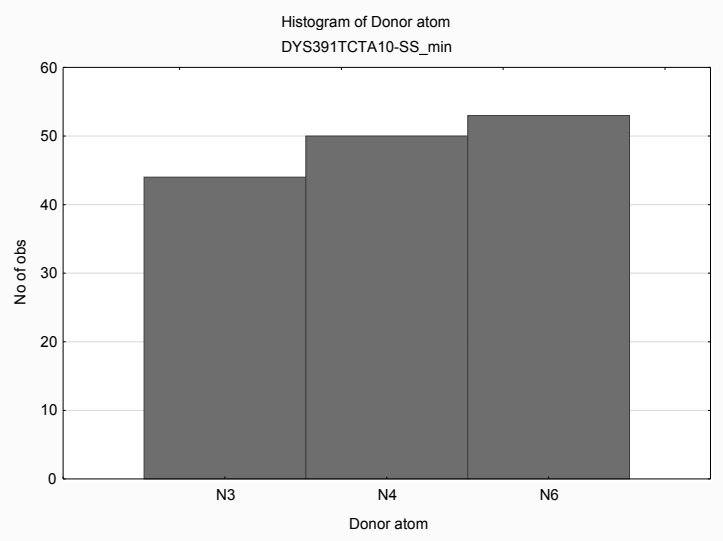

Histogram of Donor atom
DYS19A-4b13b-5b12b-UF
No of obs
0
20
40
60
80
100
120
N2
N6
N3
N1
Donor atom

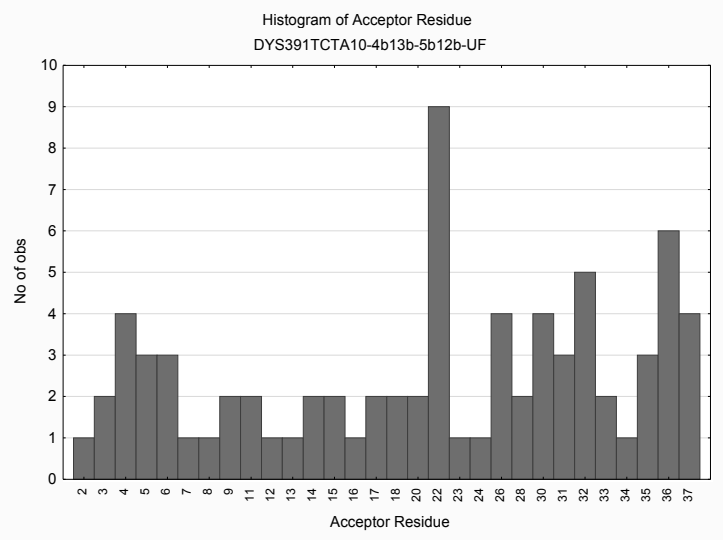
Graphs SI96-SI103: Histograms for hydrogen bonding residues and atoms (acceptors and donors) of DYS391 single-stranded and UNAFold predicted structure. Histogram of atoms considers the total number of H-bonds of the macromolecular model for each atom type. Histogram of residues considers the number of H-bonds of the macromolecular model for each residue.

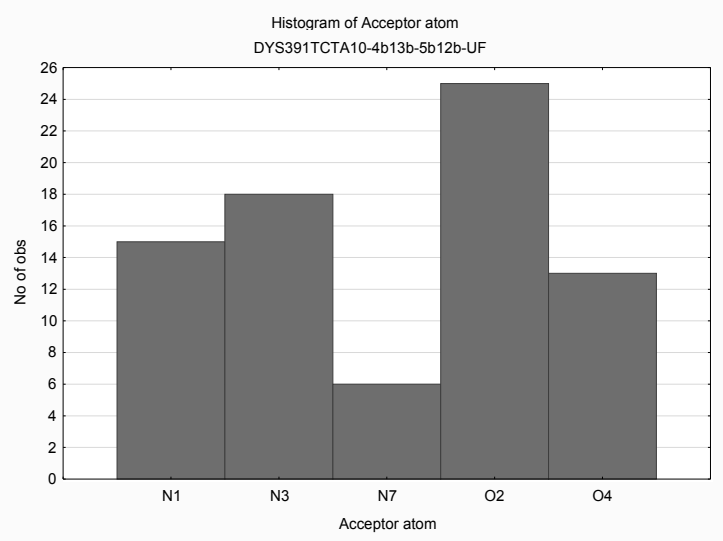


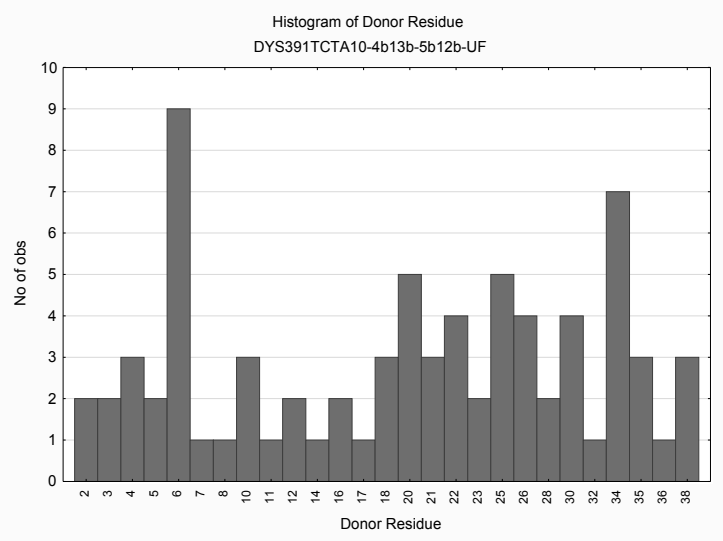


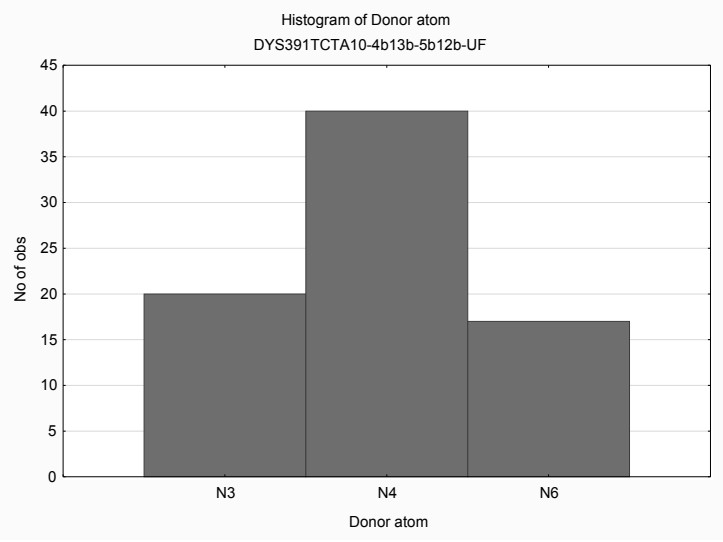

Histogram of Donor Residue
DYS391TCTA10-SS_min
No of obs
Donor Residue


Histogram of Donor atom
DYS391TCTA10-SS_min
No of obs
N3
N4
N6
Donor atom

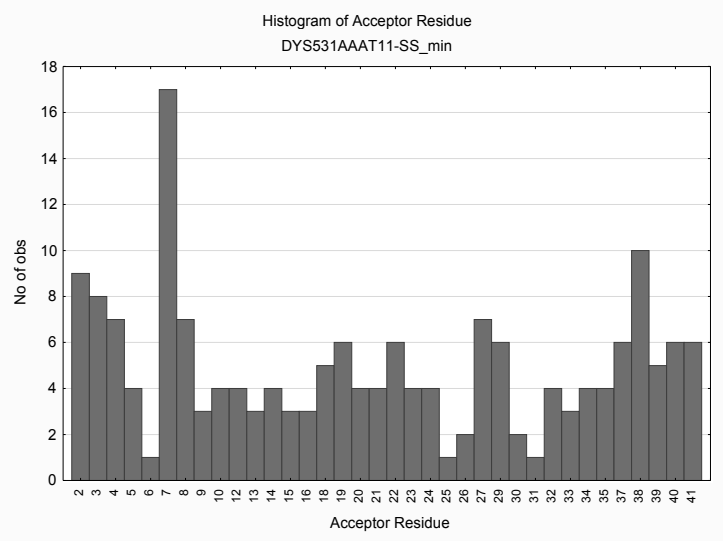

Histogram of Acceptor Residue
DYS391TCTA10-4b13b-5b12b-UF
No of obs
Acceptor Residue

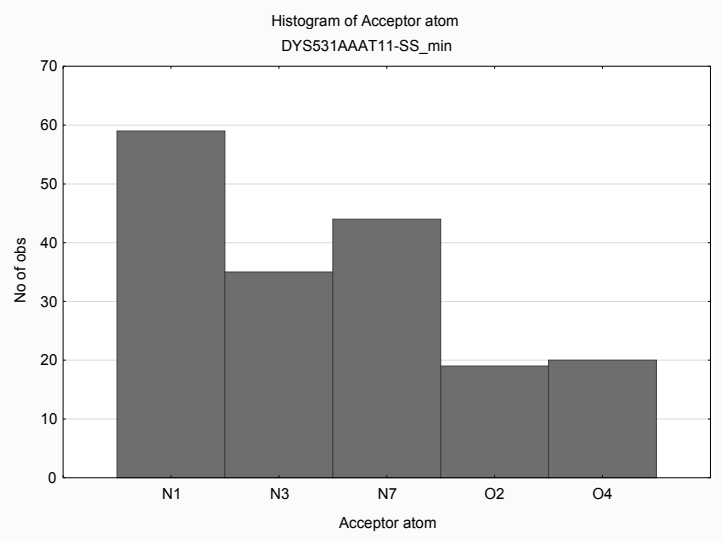

Histogram of Acceptor atom
DYS391TCTA10-4b13b-5b12b-UF
No of obs
0
2
4
6
8
10
12
14
16
18
20
22
24
26
N1
N3
N7
O2
O4
Acceptor atom

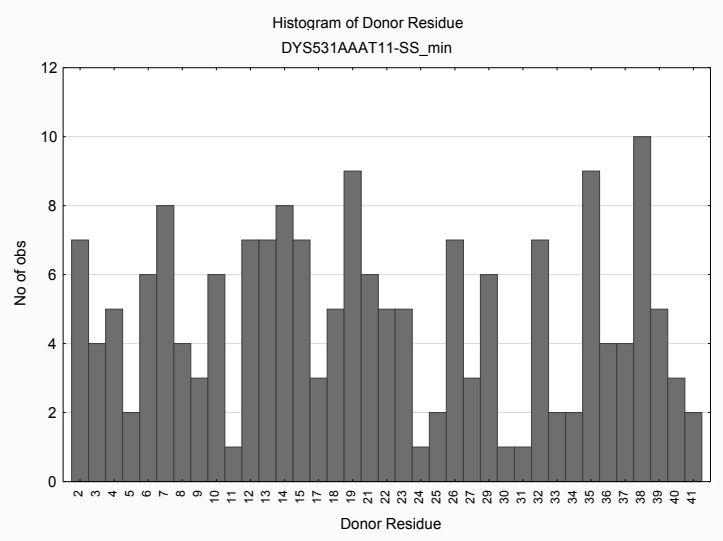

Histogram of Donor Residue
DYS391TCTA10-4b13b-5b12b-UF
No of obs
0
1
2
3
4
5
6
7
8
9
10
2
3
4
5
6
7
8
10
11
12
14
16
17
18
20
21
22
23
25
26
28
30
32
34
35
36
38
Donor Residue

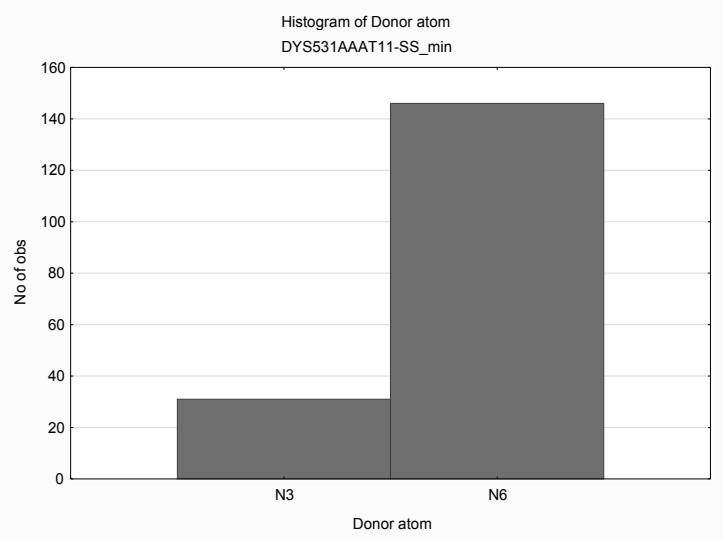

Histogram of Donor atom
DYS391TCTA10-4b13b-5b12b-UF
No of obs
0
5
10
15
20
25
30
35
40
45
N3
N4
N6
Donor atom

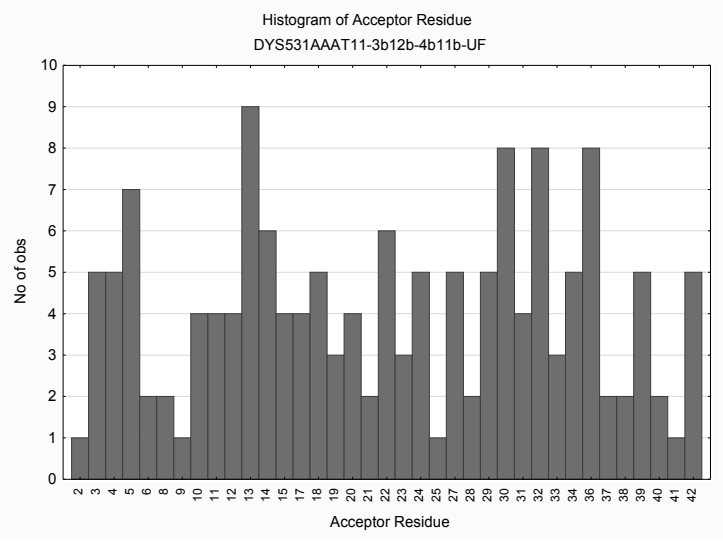
Graphs SI104-SI111: Histograms for hydrogen bonding residues and atoms (acceptors and donors) of DYS531 single-stranded and UNAFold predicted structure. Histogram of atoms considers the total number of H-bonds of the macromolecular model for each atom type. Histogram of residues considers the number of H-bonds of the macromolecular model for each residue.

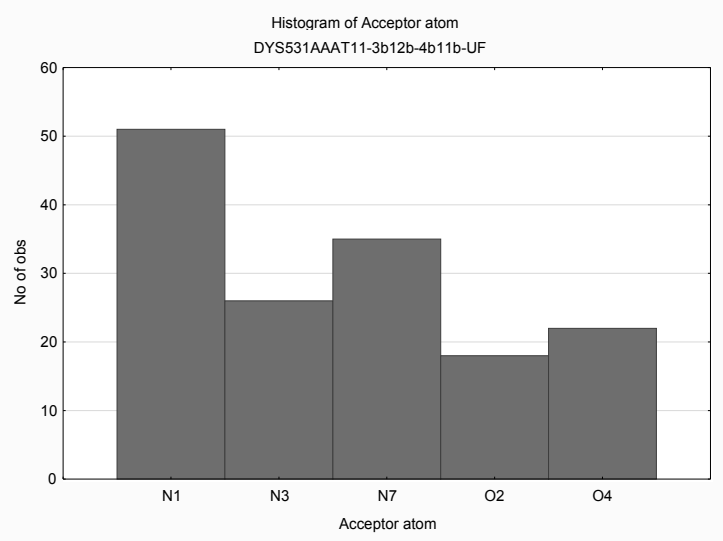


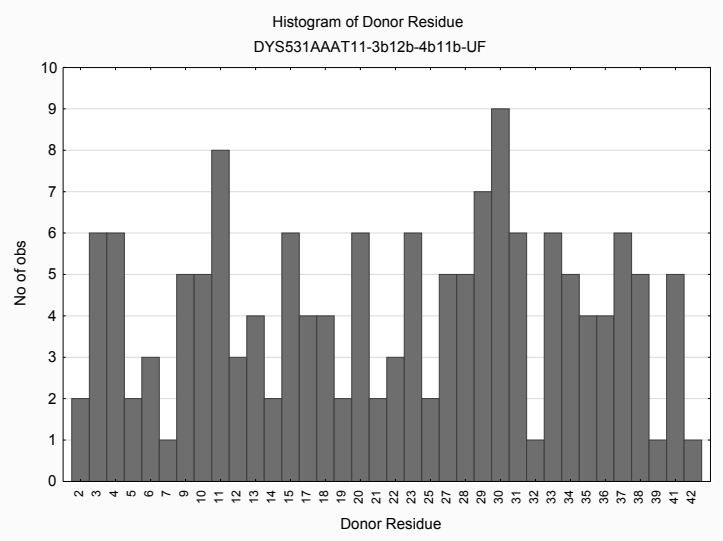


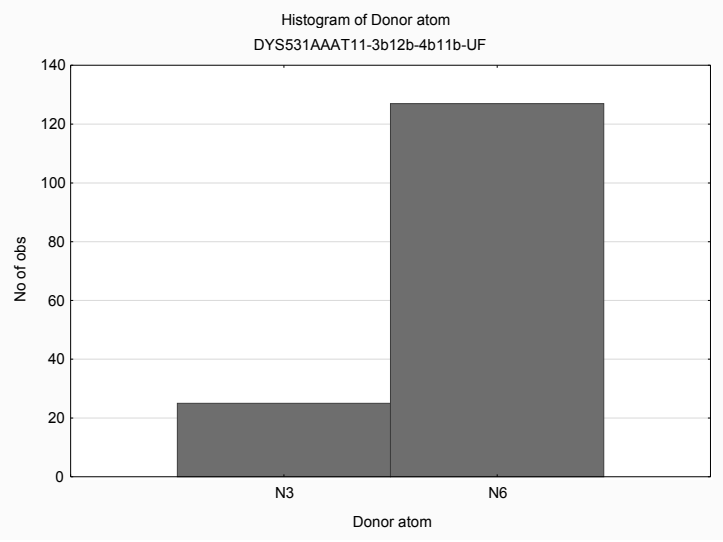

Histogram of Donor Residue
DYS531AAAT11-SS_min
No of obs
Donor Residue


Histogram of Donor atom
DYS531AAAT11-SS_min
No of obs
N3
N6
Donor atom

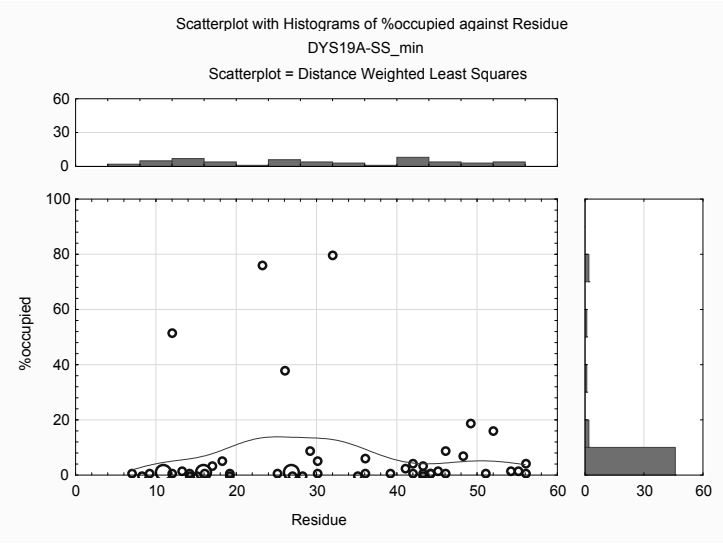

Histogram of Acceptor Residue
DYS531AAAT11-3b12b-4b11b-UF
No of obs
Acceptor Residue

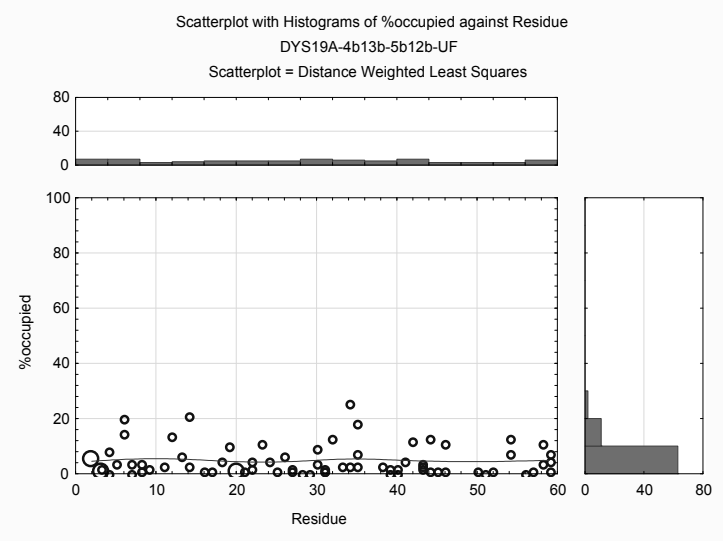

Histogram of Acceptor atom
DYS531AAAT11-3b12b-4b11b-UF
No of obs
0
10
20
30
40
50
60
N1
N3
N7
O2
O4
Acceptor atom

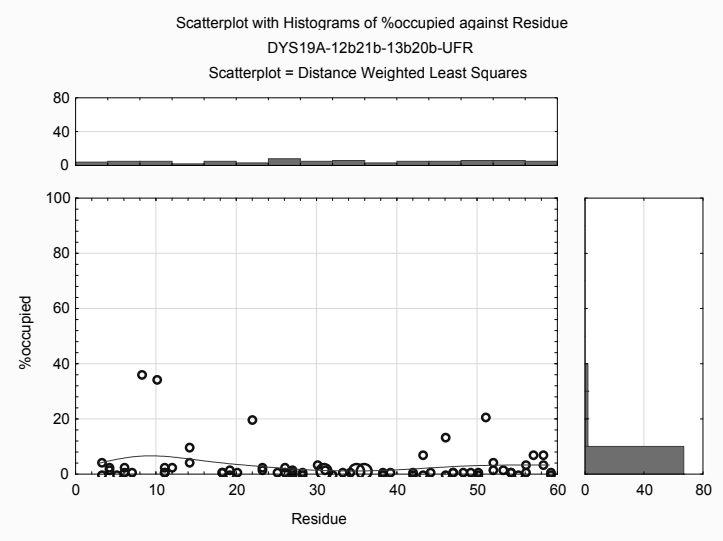

Histogram of Donor Residue
DYS531AAAT11-3b12b-4b11b-UF
No of obs
0
1
2
3
4
5
6
7
8
9
10
2
3
4
5
6
7
9
10
11
12
13
14
15
17
18
19
20
21
22
23
25
27
28
29
30
31
32
33
34
35
36
37
38
39
41
42
Donor Residue

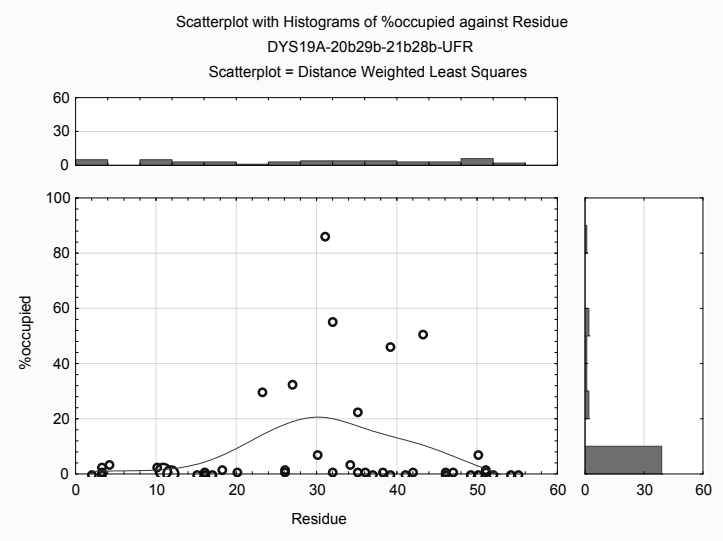

Histogram of Donor atom
DYS531AAAT11-3b12b-4b11b-UF
No of obs
0
20
40
60
80
100
120
140
N3
N6
Donor atom

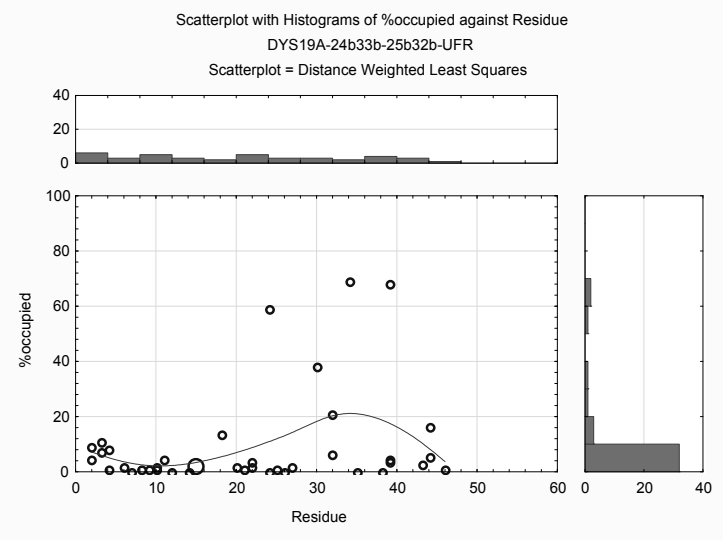
Graphs SI112-SI122: Scatterplots of counterions occupancy (Na+; cut-off 5 Å) near nucleotides (Residue position) for the last 4 ns of molecular dynamics of DYS19A tested models. A distance weighted least squares function is used to screen for proximity Na+ patterns in specific regions of DYS19A STR.

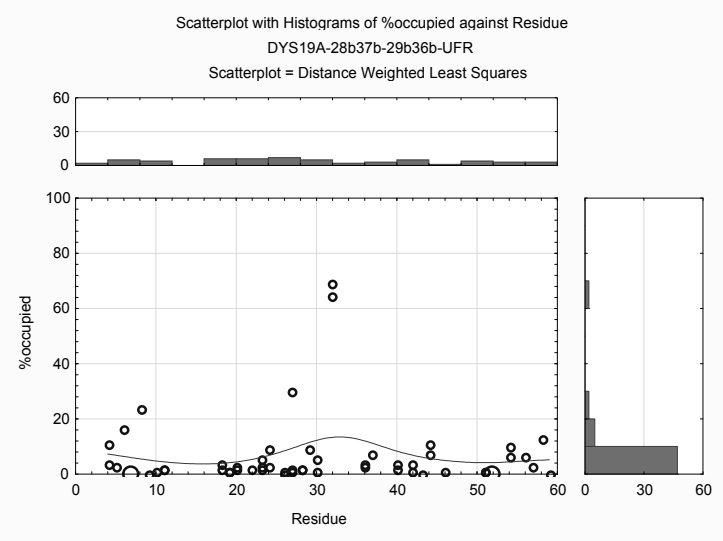


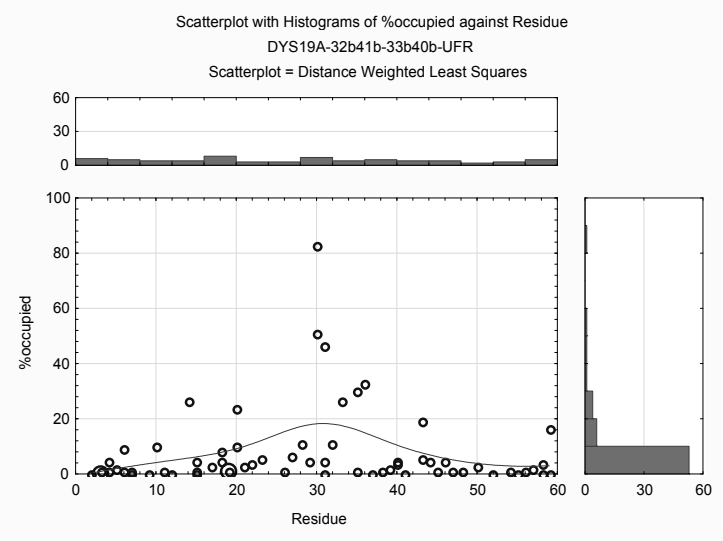


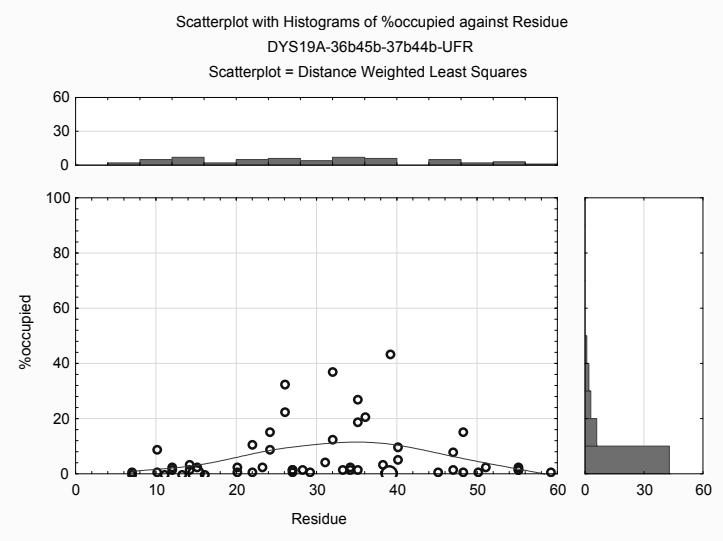

Scatterplot with Histograms of %occupied against Residue
DYS19A-12b21b-13b20b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue

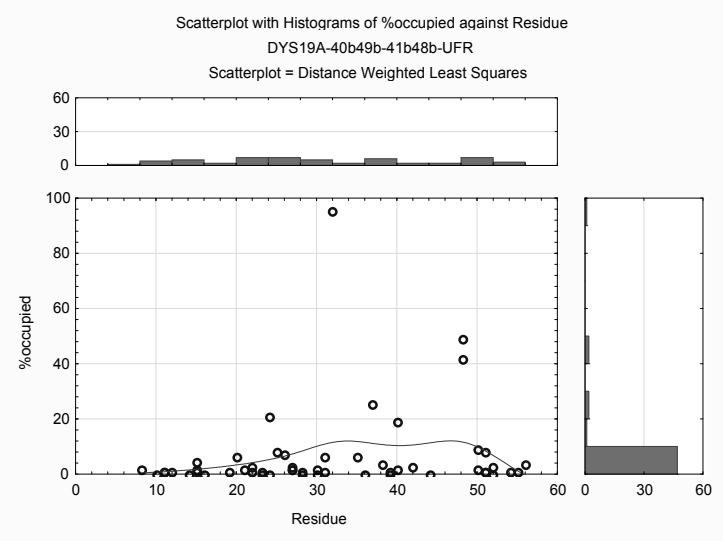

Scatterplot with Histograms of %occupied against Residue
DYS19A-20b29b-21b28b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue

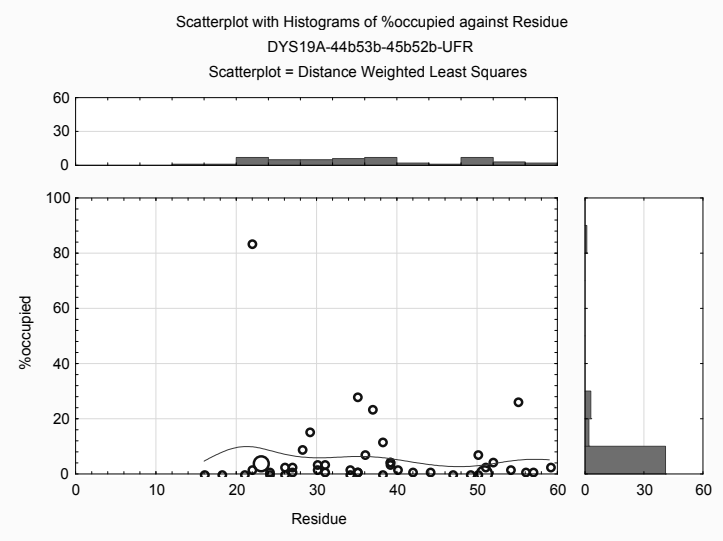

Scatterplot with Histograms of %occupied against Residue
DYS19A-24b33b-25b32b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue

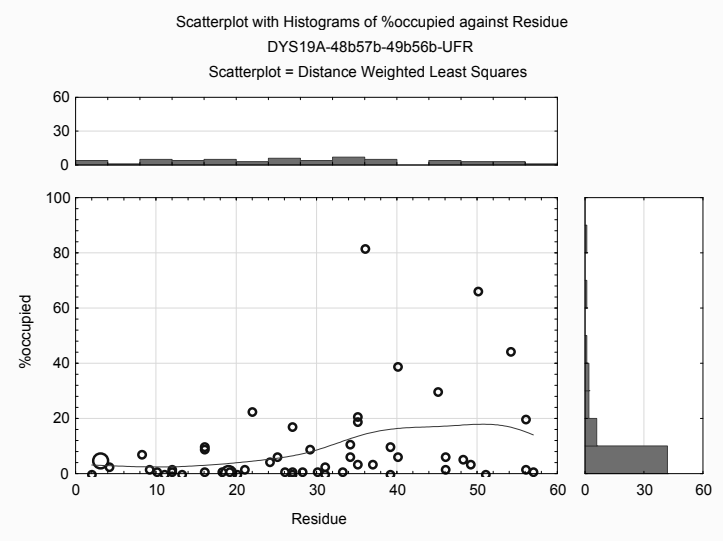

Scatterplot with Histograms of %occupied against Residue
DYS19A-28b37b-29b36b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue


Scatterplot with Histograms of %occupied against Residue
DYS19A-32b41b-33b40b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue

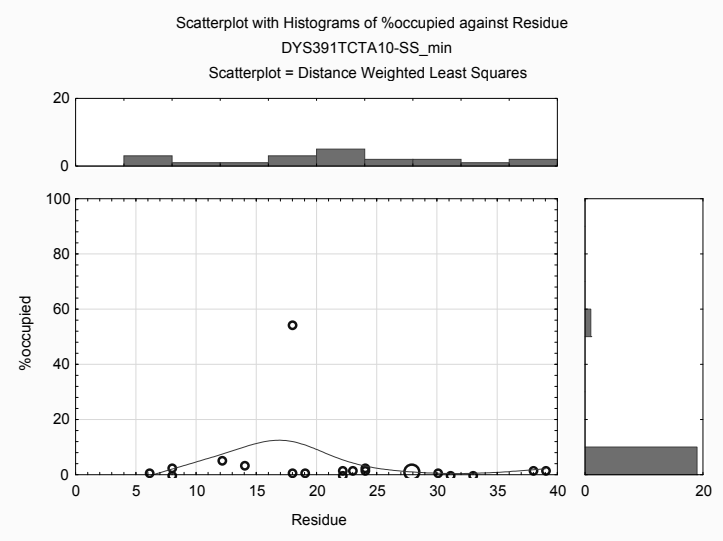

Scatterplot with Histograms of %occupied against Residue
DYS19A-36b45b-37b44b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue

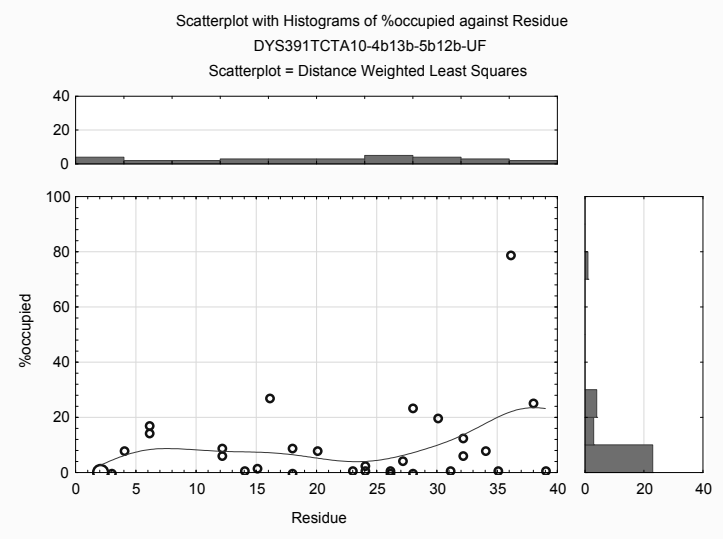

Scatterplot with Histograms of %occupied against Residue
DYS19A-40b49b-41b48b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue

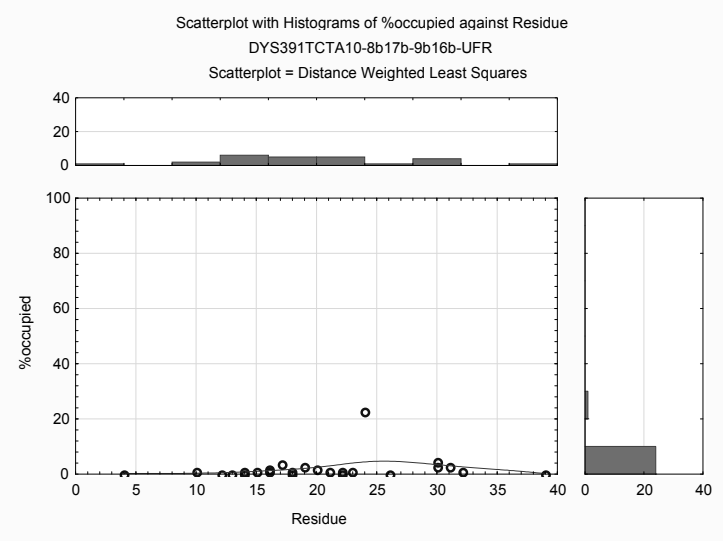

Scatterplot with Histograms of %occupied against Residue
DYS19A-44b53b-45b52b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue

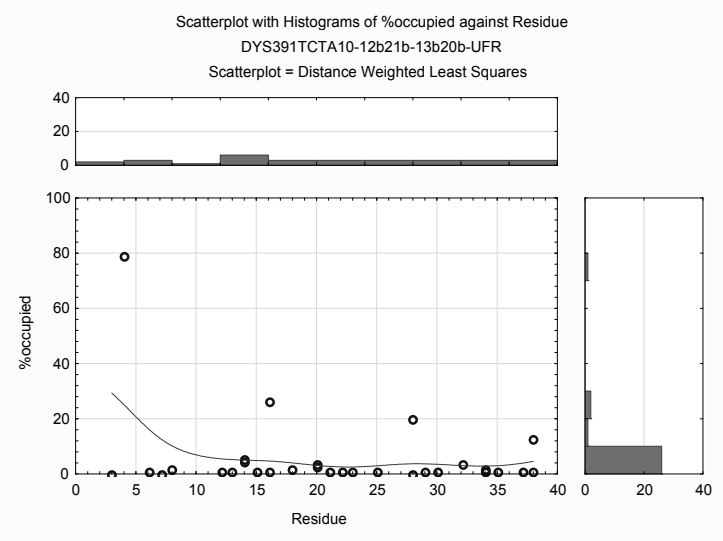

Scatterplot with Histograms of %occupied against Residue
DYS19A-48b57b-49b56b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue

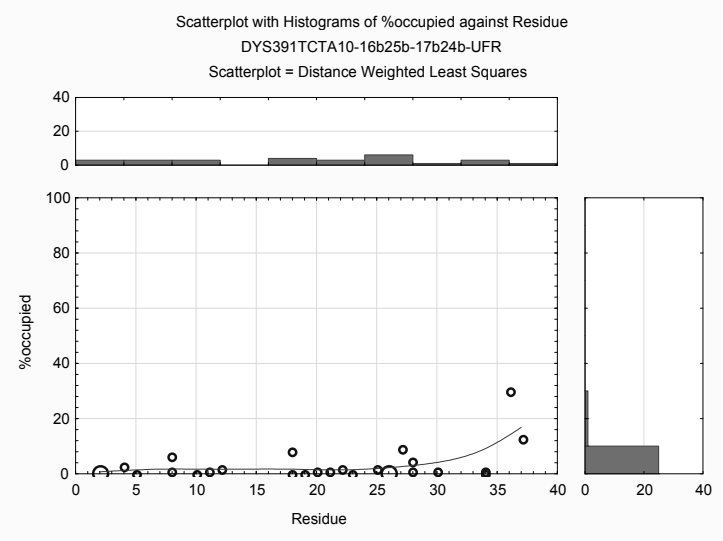
Graphs SI123-SI130: Scatterplots of ions occupancy (Na+; cut-off 5 Å) near nucleotides (Residue position) for the last 4 ns of molecular dynamics of DYS391 tested models. A distance weighted least squares function is used to screen for proximity Na+ patterns in specific regions of DYS391 STR.

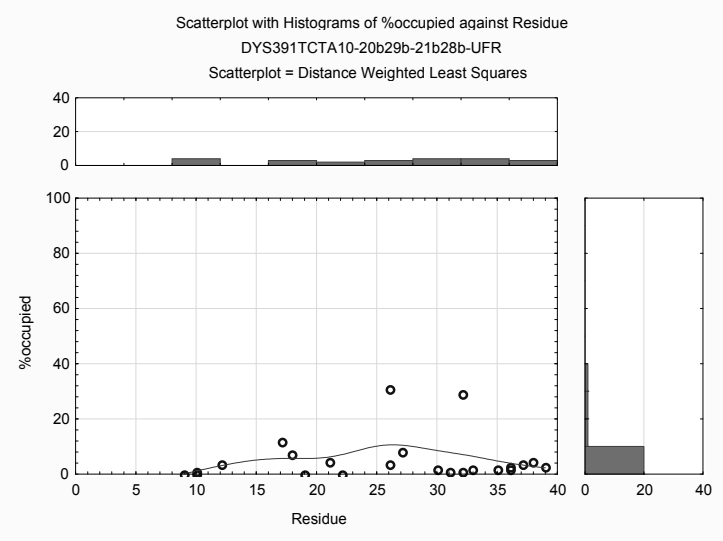


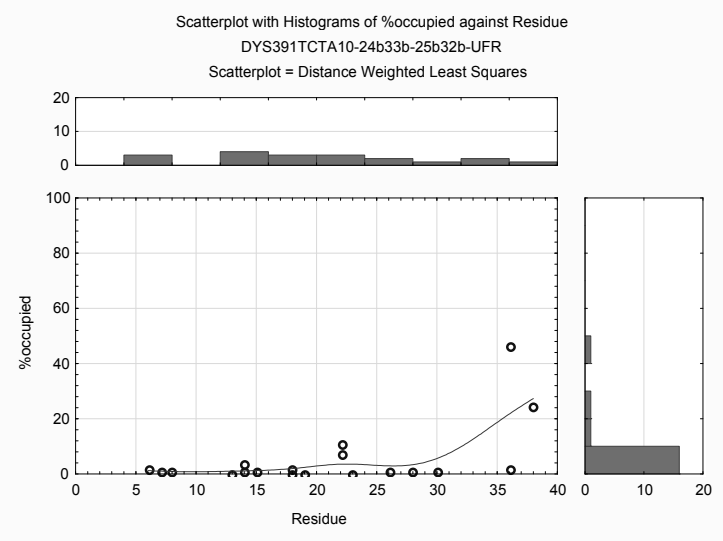


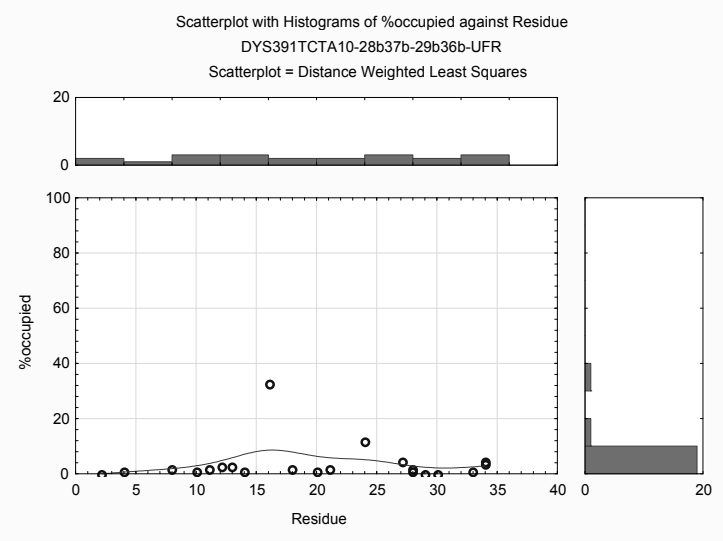

Scatterplot with Histograms of %occupied against Residue
DYS391TCTA10-8b17b-9b16b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue


Scatterplot with Histograms of %occupied against Residue
DYS391TCTA10-12b21b-13b20b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue

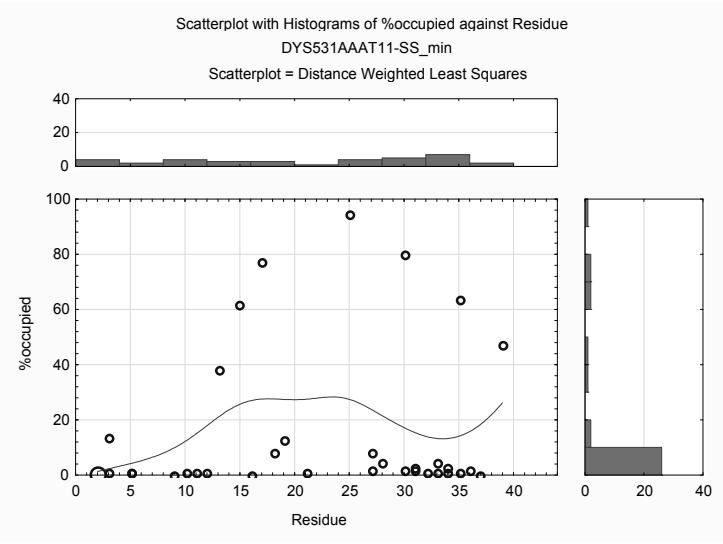

Scatterplot with Histograms of %occupied against Residue
DYS391TCTA10-16b25b-17b24b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue

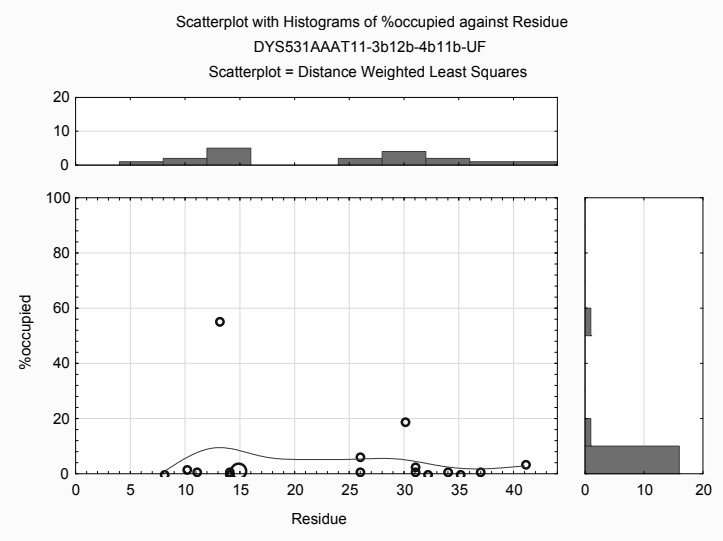

Scatterplot with Histograms of %occupied against Residue
DYS391TCTA10-20b29b-21b28b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue

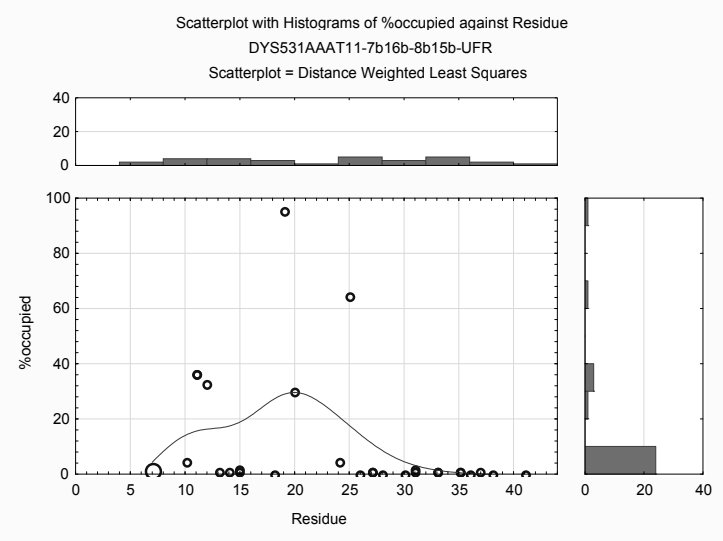

Scatterplot with Histograms of %occupied against Residue
DYS391TCTA10-24b33b-25b32b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue

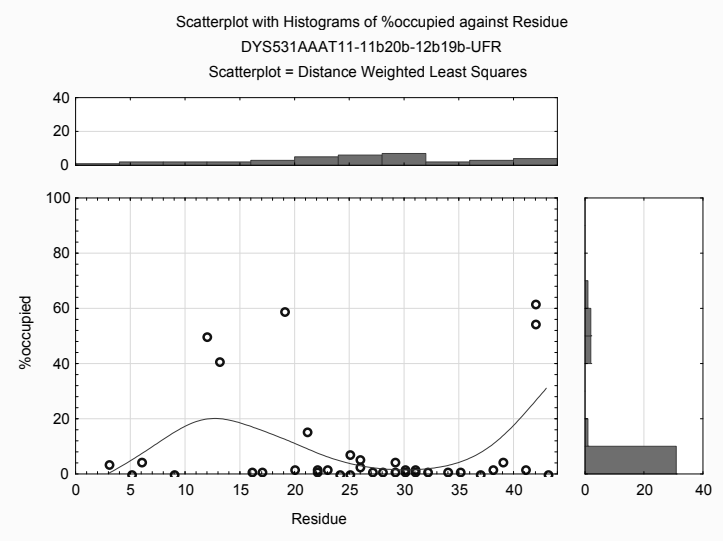

Scatterplot with Histograms of %occupied against Residue
DYS391TCTA10-28b37b-29b36b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue

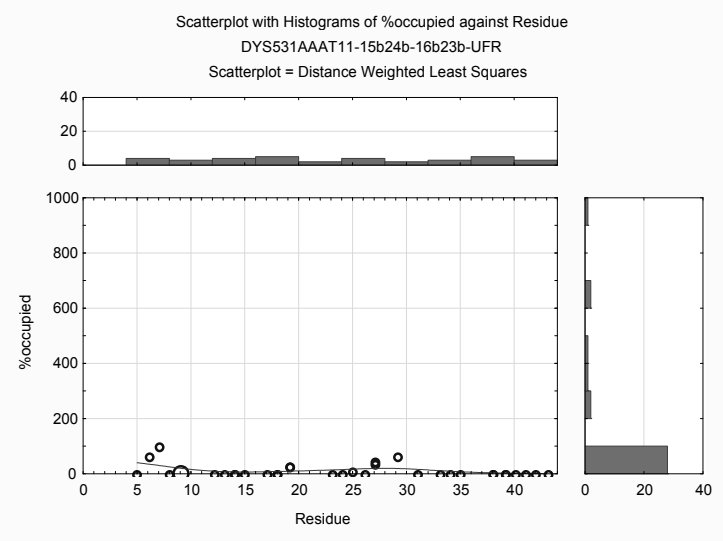
Graphs SI131-SI140: Scatterplots of ions occupancy (Na+) near nucleotides (Residue position) for the last 4 ns of molecular dynamics of DYS531 tested models. A distance weighted least squares function is used to screen for proximity Na+ patterns in specific regions of DYS531 STR.

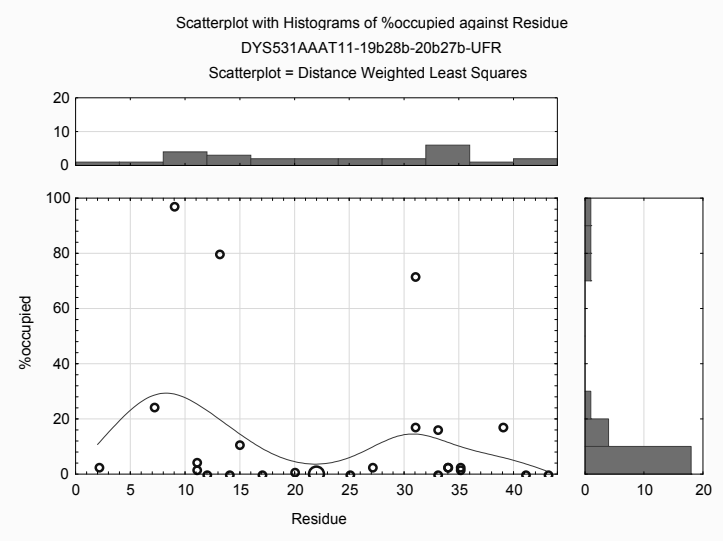


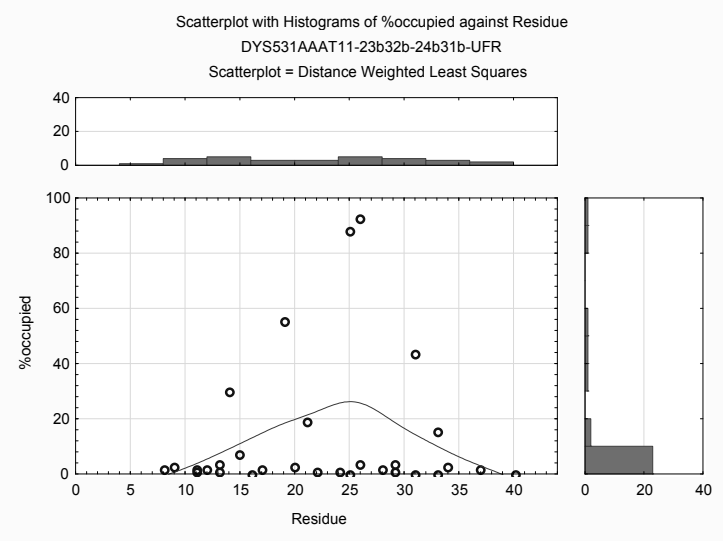


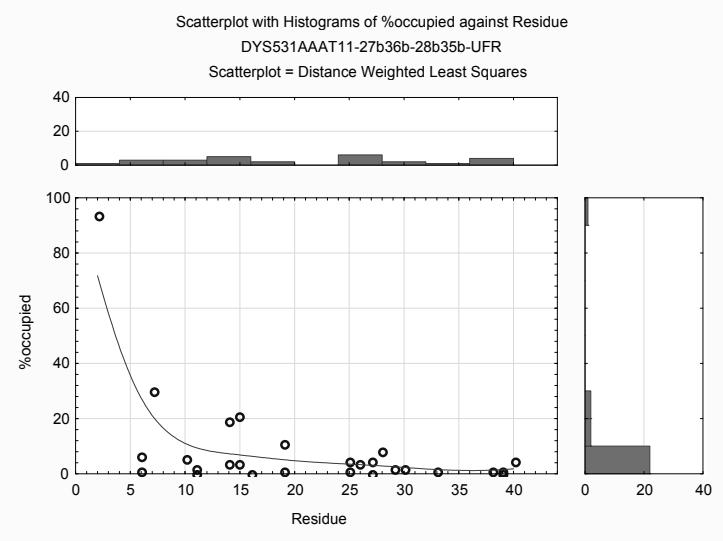

Scatterplot with Histograms of %occupied against Residue
DYS531AAAT11-7b16b-8b15b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue

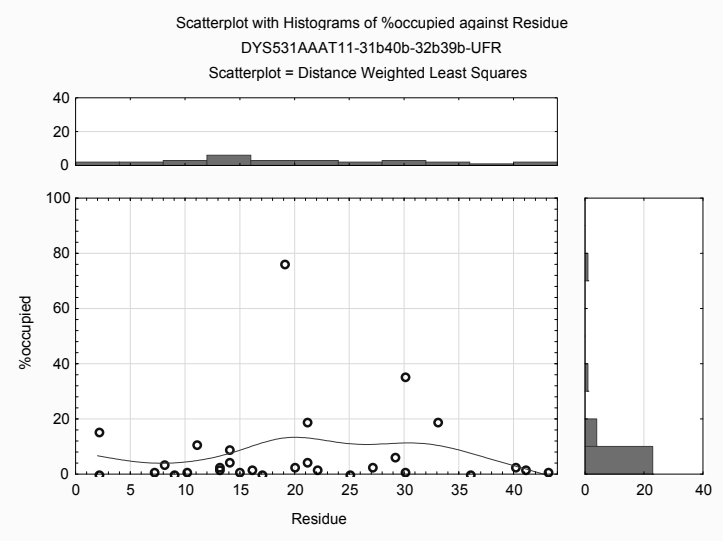

Scatterplot with Histograms of %occupied against Residue
DYS531AAAT11-11b20b-12b19b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue

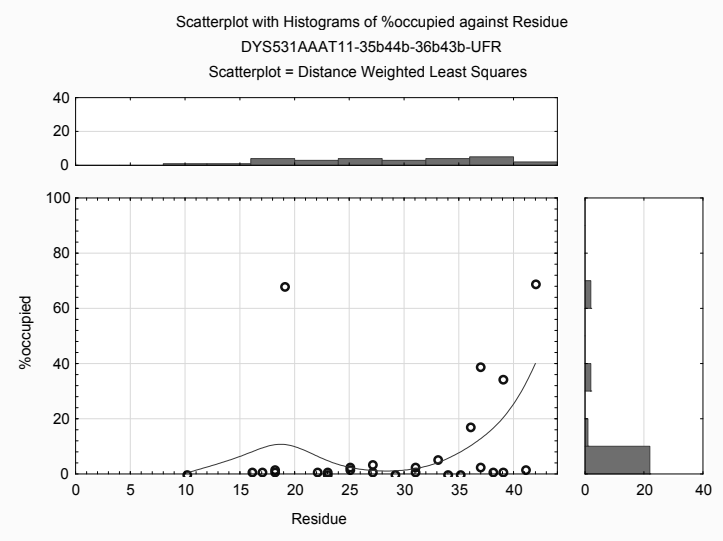

Scatterplot with Histograms of %occupied against Residue
DYS531AAAT11-15b24b-16b23b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue


Scatterplot with Histograms of %occupied against Residue
DYS531AAAT11-19b28b-20b27b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue


Scatterplot with Histograms of %occupied against Residue
DYS531AAAT11-23b32b-24b31b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue

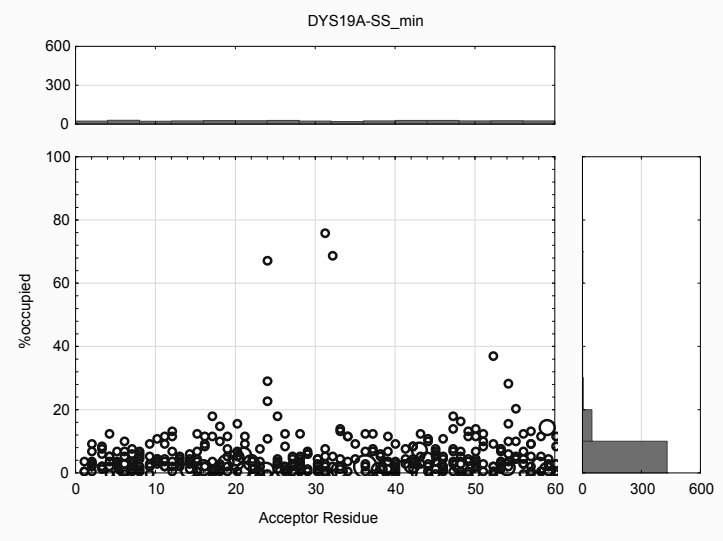

Scatterplot with Histograms of %occupied against Residue
DYS531AAAT11-27b36b-28b35b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue


Scatterplot with Histograms of %occupied against Residue
DYS531AAAT11-31b40b-32b39b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue

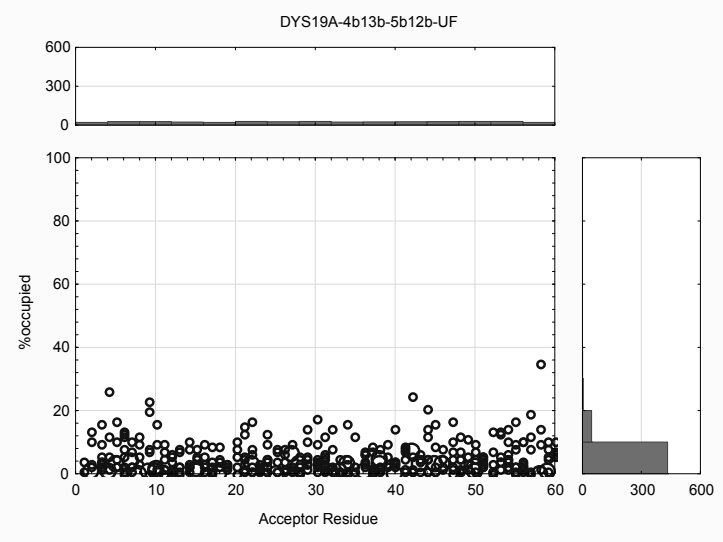

Scatterplot with Histograms of %occupied against Residue
DYS531AAAT11-35b44b-36b43b-UFR
Scatterplot = Distance Weighted Least Squares
%occupied
Residue

Graphs SI141-SI151: Scatterplots of water occupancy (WAT; cut-off 3 Å) near nucleotides (Acceptor Residue position) for the last 4 ns of molecular dynamics of DYS19A tested models. Valid N corresponds to all H-bonds detected in the last 4 ns of molecular dynamics simulation.

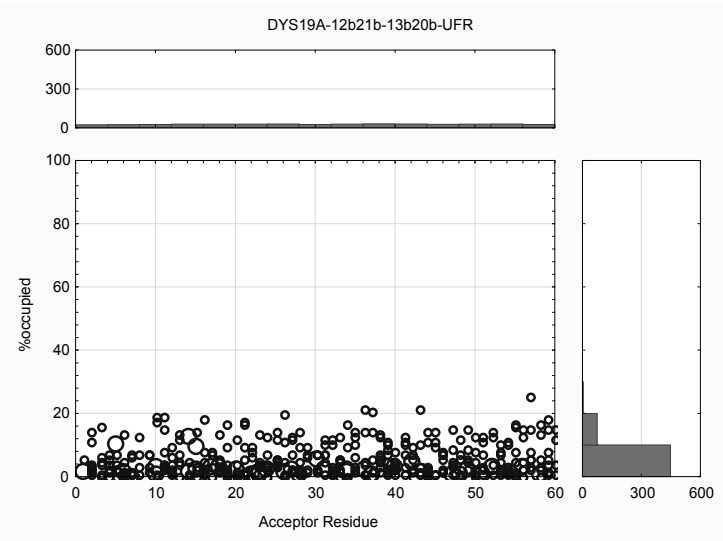
	DYS19A-SS_min				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	487	4.73	0.30	76.35	6.91
distance	487	2.81	2.66	2.99	0.07
angle	487	24.59	2.20	58.89	7.99
lifetime	487	5.21	4.00	28.30	1.96
maxocc	487	2.57	1.00	39.00	2.74



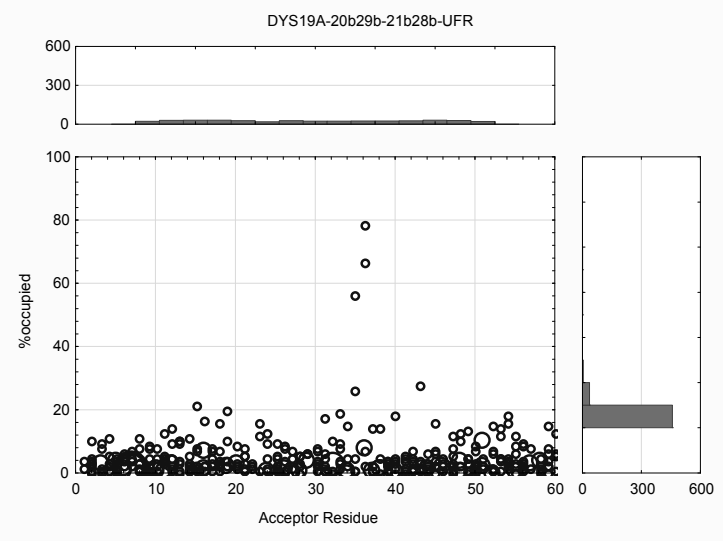
	DYS19A-4b13b-5b12b-UF				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	486	4.30	0.30	35.03	4.53
distance	486	2.82	2.60	3.00	0.07
angle	486	25.59	4.79	59.84	8.83
lifetime	486	5.43	4.00	62.70	3.25
maxocc	486	2.68	1.00	28.00	2.65



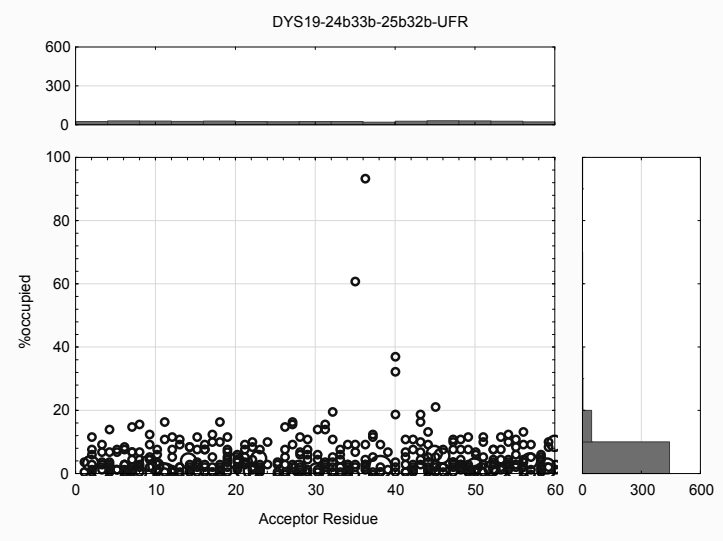
	DYS19A-12b21b-13b20b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	528	4.62	0.30	25.45	4.70
distance	528	2.82	2.63	2.98	0.07
angle	528	24.82	4.75	58.13	7.83
lifetime	528	5.27	4.00	15.70	1.67
maxocc	528	2.67	1.00	17.00	2.23



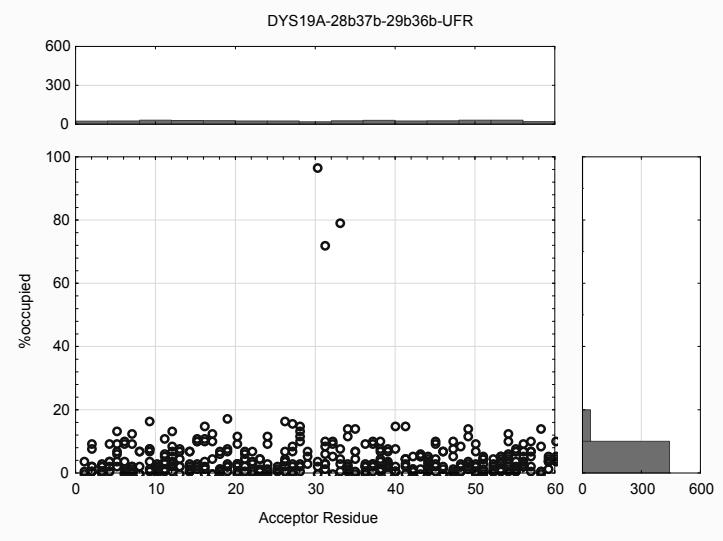
	DYS19A-20b29b-21b28b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	501	4.09	0.30	79.04	6.41
distance	501	2.82	2.59	2.98	0.07
angle	501	24.99	2.43	59.22	8.66
lifetime	501	5.09	4.00	64.00	3.08
maxocc	501	2.35	1.00	21.00	2.22



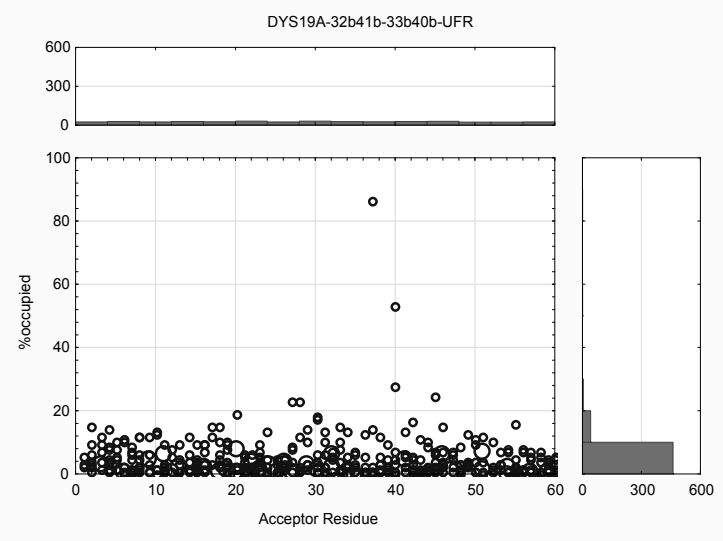
	DYS19-24b33b-25b32b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	497	4.56	0.30	93.41	6.58
distance	497	2.82	2.65	2.99	0.07
angle	497	24.51	3.50	58.77	7.78
lifetime	497	5.16	4.00	59.40	2.78
maxocc	497	2.55	1.00	63.00	3.28



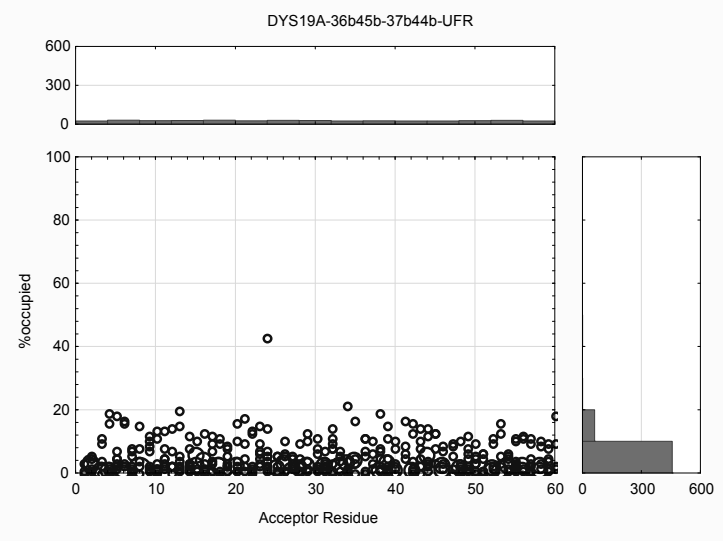
Variable	DYS19A-28b37b-29b36b-UFR				
	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	488	4.34	0.30	97.01	7.55
distance	488	2.82	2.61	3.00	0.07
angle	488	25.01	2.15	59.24	8.52
lifetime	488	5.20	4.00	81.00	3.90
maxocc	488	2.60	1.00	91.00	5.16



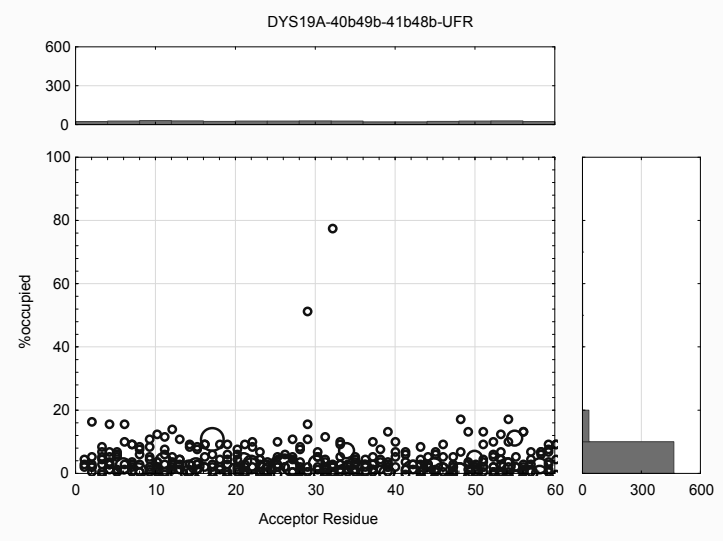
Variable	DYS19A-32b41b-33b40b-UFR				
	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	509	4.12	0.30	86.83	5.95
distance	509	2.81	2.63	2.99	0.07
angle	509	25.10	4.93	58.84	8.41
lifetime	509	5.19	4.00	21.50	1.94
maxocc	509	2.45	1.00	35.00	2.62



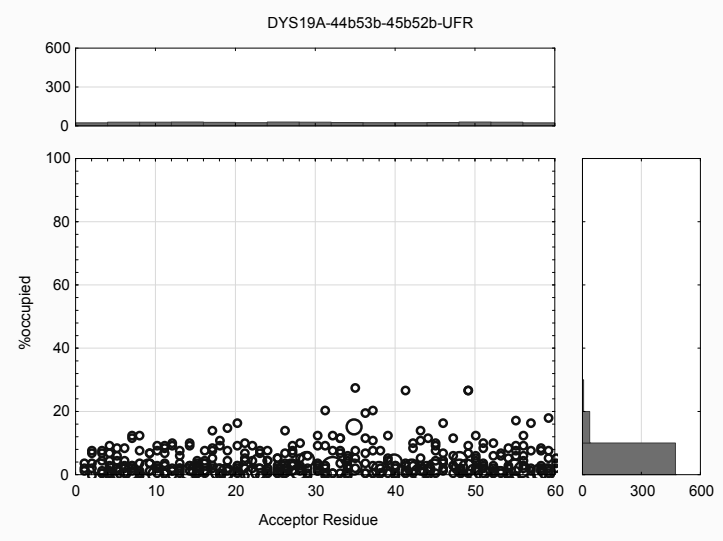
Variable	DYS19A-36b45b-37b44b-UFR				
	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	521	4.16	0.30	43.11	4.58
distance	521	2.82	2.66	3.00	0.08
angle	521	25.16	10.36	57.11	7.63
lifetime	521	5.16	4.00	19.30	1.81
maxocc	521	2.51	1.00	24.00	2.45



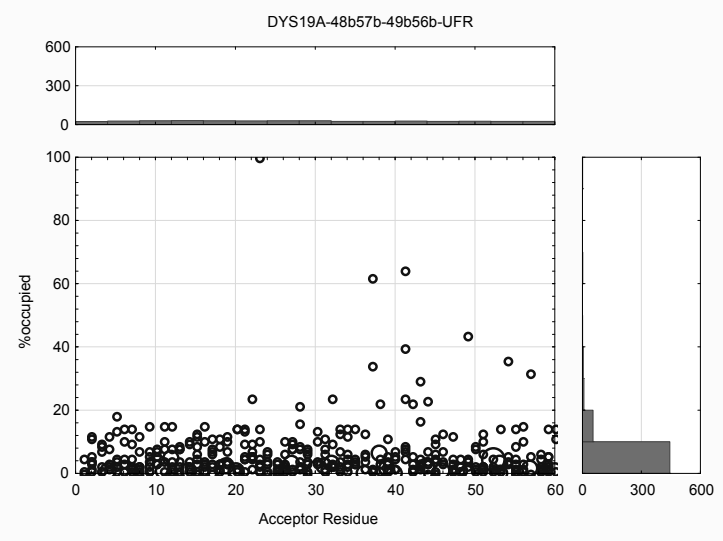
Variable	DYS19A-40b49b-41b48b-UFR				
	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	501	3.84	0.30	77.84	5.60
distance	501	2.82	2.56	2.99	0.08
angle	501	24.63	4.73	59.12	8.53
lifetime	501	5.27	4.00	26.00	2.22
maxocc	501	2.53	1.00	53.00	3.08



Variable	DYS19A-44b53b-45b52b-UFR				
	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	516	3.90	0.30	27.54	4.34
distance	516	2.82	2.63	3.00	0.07
angle	516	25.30	3.19	58.35	8.40
lifetime	516	5.19	4.00	20.00	1.86
maxocc	516	2.44	1.00	19.00	2.14

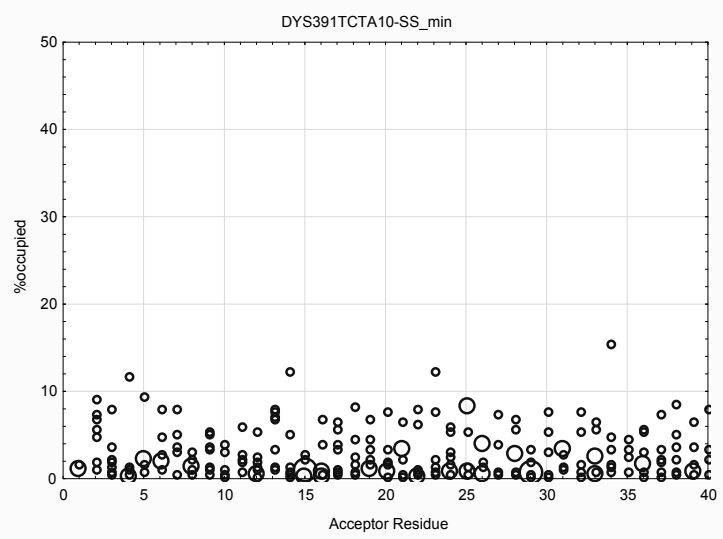


Variable	DYS19A-48b57b-49b56b-UFR				
	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	515	5.11	0.30	100.00	7.85
distance	515	2.82	2.57	2.99	0.07
angle	515	25.16	4.58	59.62	8.16
lifetime	515	5.29	4.00	29.60	2.73
maxocc	515	2.83	1.00	54.00	4.46

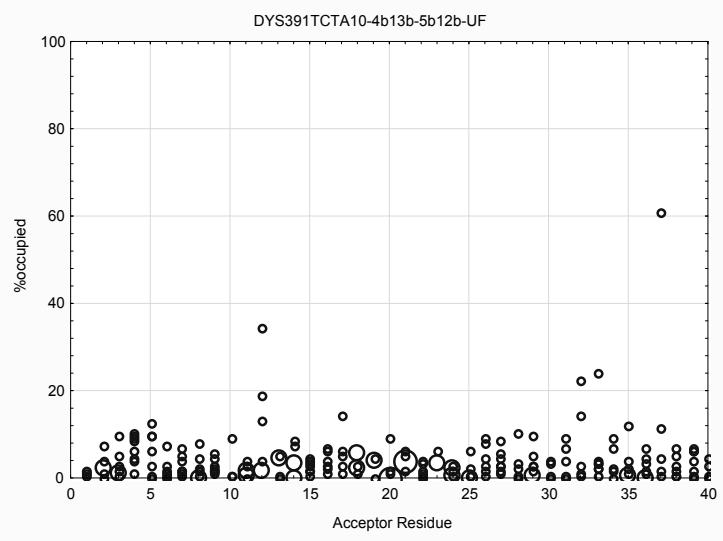


Graphs SI152-SI159: Scatterplots of water occupancy (WAT; cut-off 3 Å) near nucleotides (Acceptor Residue position) for the last 4 ns of molecular dynamics of DYS391 tested models. Valid N corresponds to all H-bonds detected in the last 4 ns of molecular dynamics simulation.

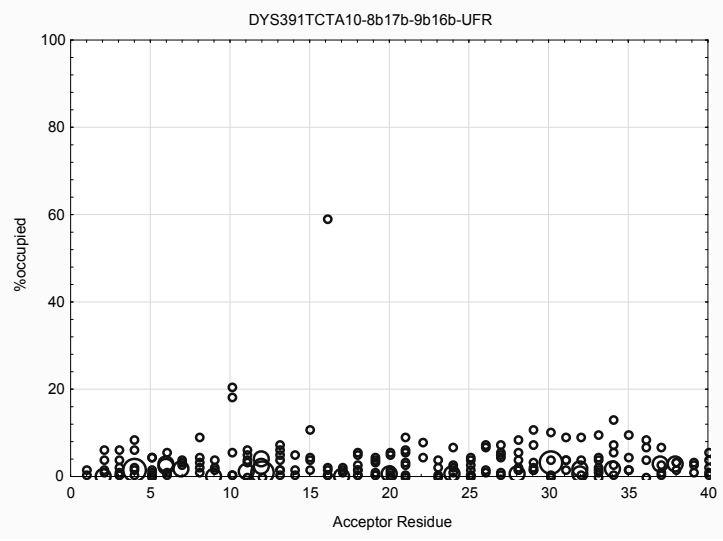
	DYS391TCTA10-SS_min				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	290	2.66	0.30	15.57	2.61
distance	290	2.80	2.59	2.99	0.08
angle	290	24.38	7.06	52.58	7.97
lifetime	290	5.01	4.00	13.00	1.52
maxocc	290	2.13	1.00	11.00	1.77



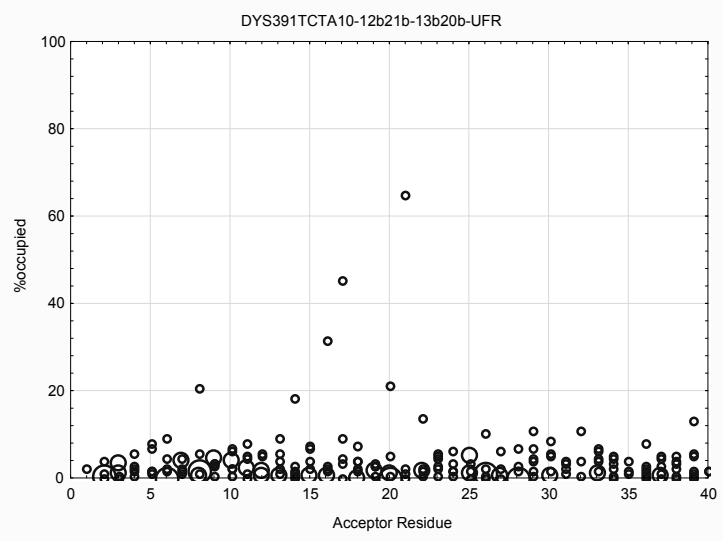
Variable	DYS391TCTA10-4b13b-5b12b-UF				
	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	268	3.69	0.30	61.08	5.42
distance	268	2.81	2.65	2.99	0.08
angle	268	24.61	4.10	56.68	9.15
lifetime	268	5.90	4.00	78.00	5.51
maxocc	268	2.96	1.00	47.00	4.29



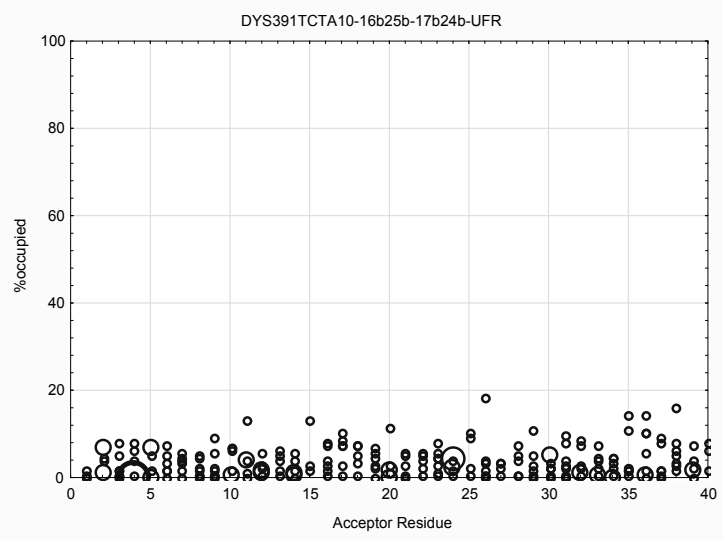
Variable	DYS391TCTA10-8b17b-9b16b-UFR				
	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	272	3.05	0.30	58.98	4.45
distance	272	2.80	2.55	2.99	0.08
angle	272	24.23	8.93	58.03	8.61
lifetime	272	5.47	4.00	49.20	3.69
maxocc	272	2.49	1.00	50.00	3.54



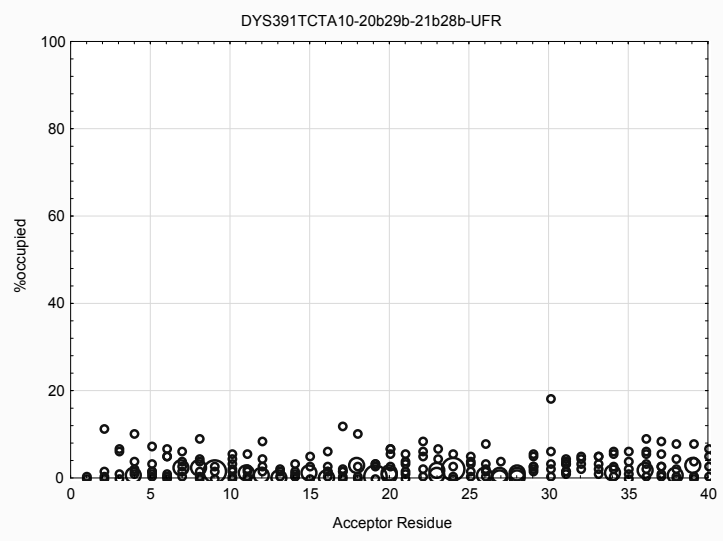
Variable	DYS391TCTA10-12b21b-13b20b-UFR				
	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	275	3.60	0.30	108.38	8.56
distance	275	2.80	2.59	3.00	0.08
angle	275	25.02	2.04	58.93	9.38
lifetime	275	5.49	4.00	53.60	3.62
maxocc	275	2.94	1.00	159.00	9.73



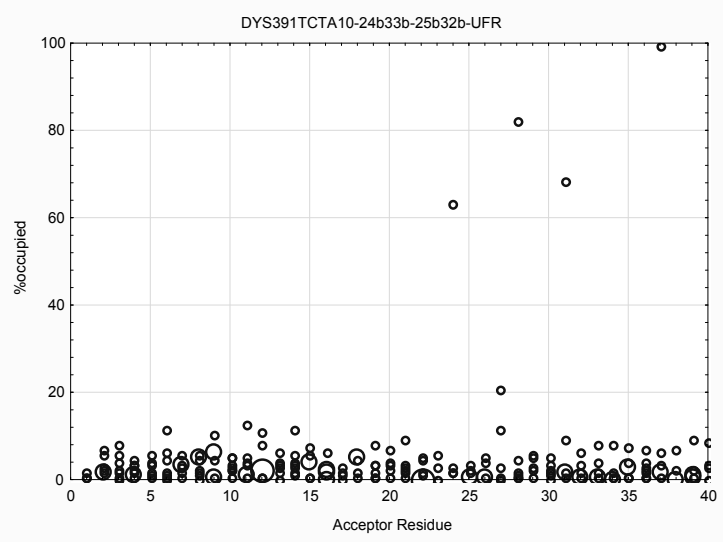
Variable	DYS391TCTA10-16b25b-17b24b-UFR				
	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	292	3.21	0.30	18.56	3.21
distance	292	2.80	2.49	3.00	0.08
angle	292	23.81	3.09	57.36	8.74
lifetime	292	5.41	4.00	24.00	2.46
maxocc	292	2.62	1.00	25.00	2.76



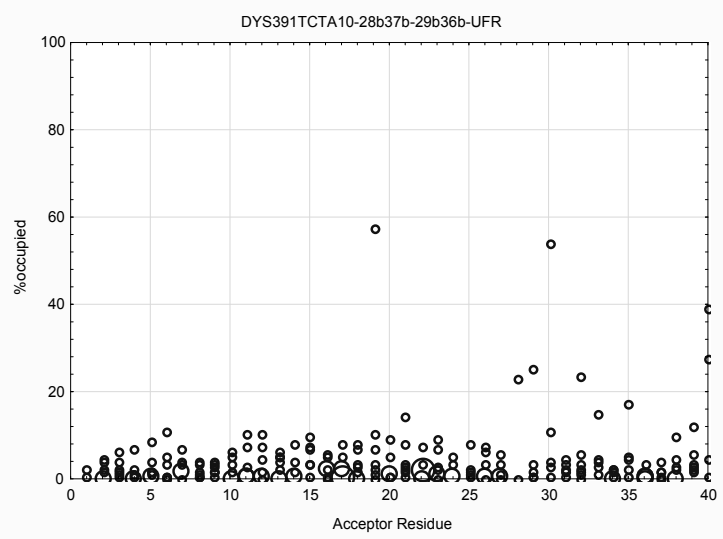
Variable	DYS391TCTA10-20b29b-21b28b-UFR				
	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	281	2.53	0.30	18.26	2.53
distance	281	2.81	2.60	2.99	0.08
angle	281	25.19	2.56	57.52	8.72
lifetime	281	5.46	4.00	28.00	2.42
maxocc	281	2.40	1.00	16.00	2.34



Variable	DYS391TCTA10-24b33b-25b32b-UFR				
	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	282	3.77	0.30	99.40	9.50
distance	282	2.81	2.63	2.99	0.07
angle	282	25.26	4.72	58.57	9.16
lifetime	282	6.80	4.00	442.70	26.17
maxocc	282	2.87	1.00	185.00	11.04

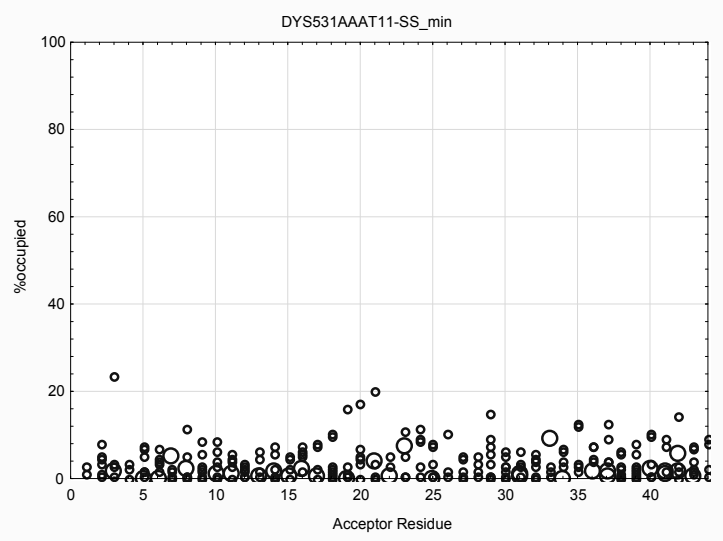


	DYS391TCTA10-28b37b-29b36b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
maxocc	267	2.58	1.00	25.00	2.78
lifetime	267	5.40	4.00	28.60	2.42
angle	267	24.94	5.09	58.39	8.93
distance	267	2.81	2.63	3.00	0.08
%occupied	267	3.78	0.30	57.49	6.62

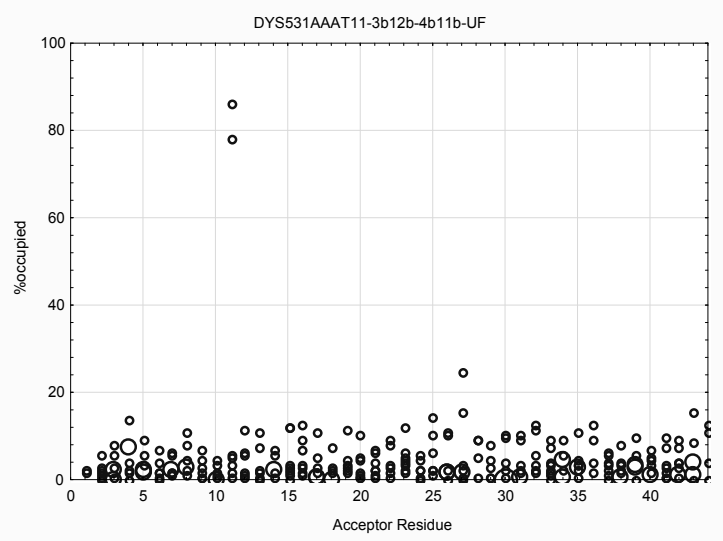


Graphs SI160-SI169: Scatterplots of water occupancy (WAT; cut-off 3 Å) near nucleotides (Acceptor Residue position) for the last 4 ns of molecular dynamics of DYS531 tested models. Valid N corresponds to all H-bonds detected in the last 4 ns of molecular dynamics simulation.

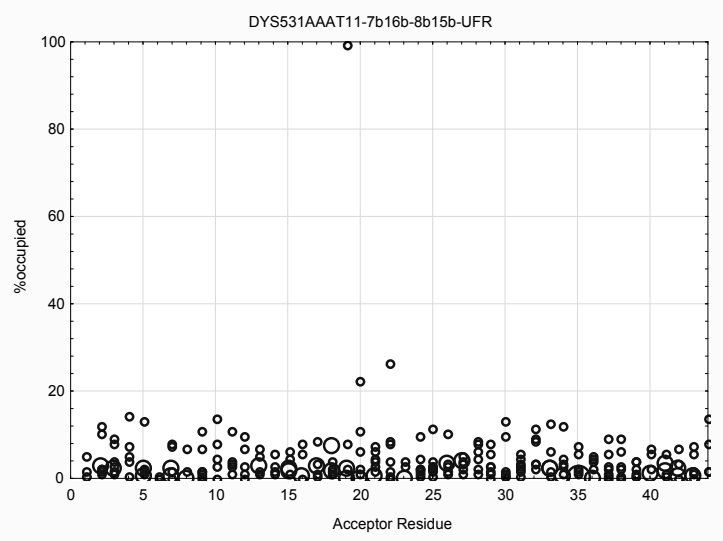
	DYS531AAAT11-SS_min				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	327	3.26	0.30	23.35	3.51
distance	327	2.82	2.67	3.00	0.08
angle	327	25.17	6.11	59.12	9.25
lifetime	327	5.13	4.00	20.80	1.93
maxocc	327	2.28	1.00	15.00	2.00



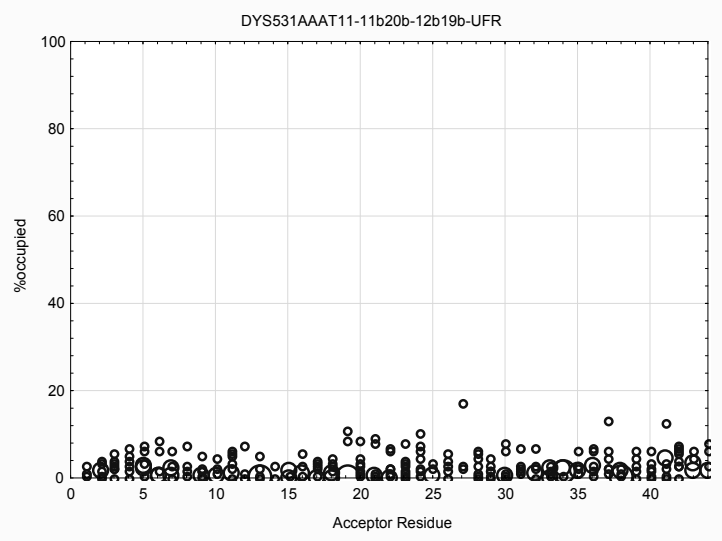
	DYS531AAAT11-3b12b-4b11b-UF				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	360	3.94	0.30	86.23	6.84
distance	360	2.82	2.52	3.00	0.08
angle	360	24.38	3.76	54.38	7.90
lifetime	360	5.17	4.00	23.50	2.12
maxocc	360	2.53	1.00	33.00	3.05



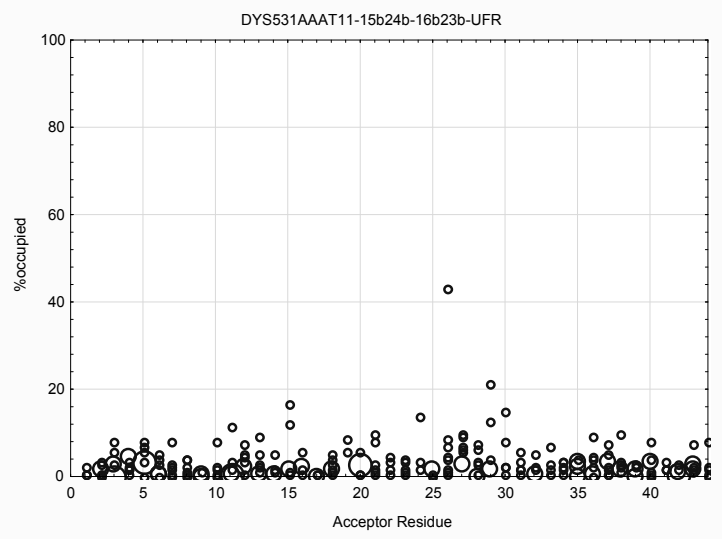
	DYS531AAAT11-7b16b-8b15b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	336	3.49	0.30	99.40	6.31
distance	336	2.82	2.60	3.00	0.08
angle	336	24.35	7.11	57.76	8.31
lifetime	336	5.05	4.00	34.10	2.15
maxocc	336	2.47	1.00	101.00	5.66



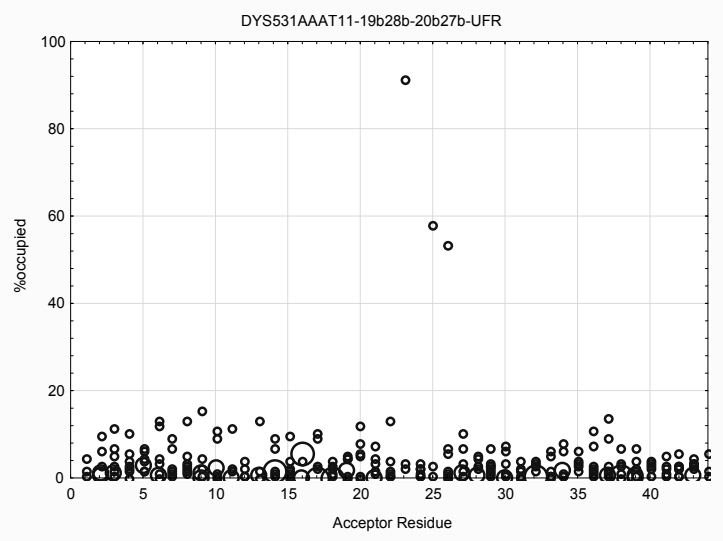
	DYS531AAAT11-11b20b-12b19b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	333	2.45	0.30	17.07	2.50
distance	333	2.82	2.65	3.00	0.08
angle	333	24.73	1.49	57.71	9.31
lifetime	333	4.96	4.00	20.70	1.72
maxocc	333	2.05	1.00	12.00	1.76



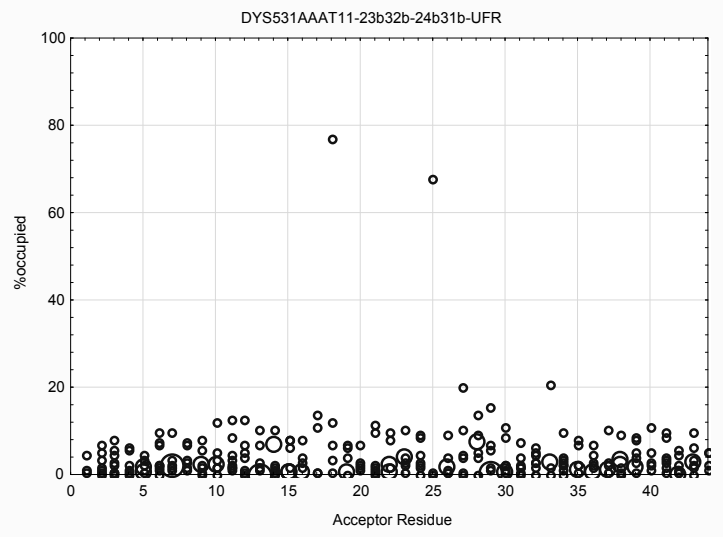
	DYS531AAAT11-15b24b-16b23b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	307	2.83	0.30	43.41	3.74
distance	307	2.81	2.62	3.00	0.08
angle	307	25.12	3.32	59.43	9.05
lifetime	307	4.93	4.00	18.00	1.62
maxocc	307	2.04	1.00	15.00	1.80



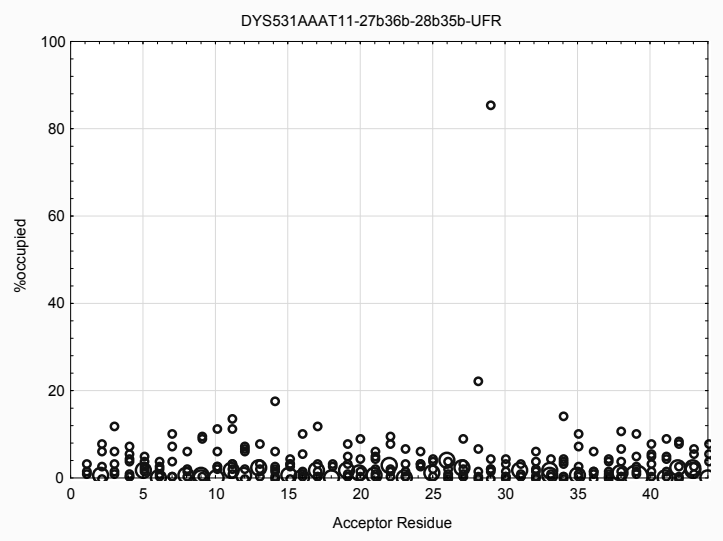
	DYS531AAAT11-19b28b-20b27b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	337	3.34	0.30	91.62	6.98
distance	337	2.81	2.64	3.00	0.08
angle	337	25.41	2.20	59.79	9.68
lifetime	337	5.13	4.00	36.00	2.32
maxocc	337	2.65	1.00	108.00	6.50



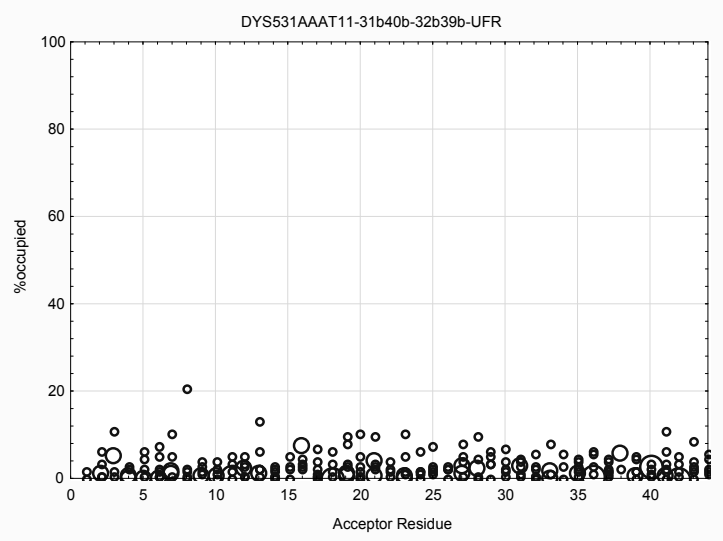
	DYS531AAAT11-23b32b-24b31b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	351	3.63	0.30	76.95	6.24
distance	351	2.81	2.50	2.99	0.08
angle	351	24.10	4.19	55.71	8.60
lifetime	351	5.25	4.00	36.70	2.86
maxocc	351	2.48	1.00	43.00	3.30



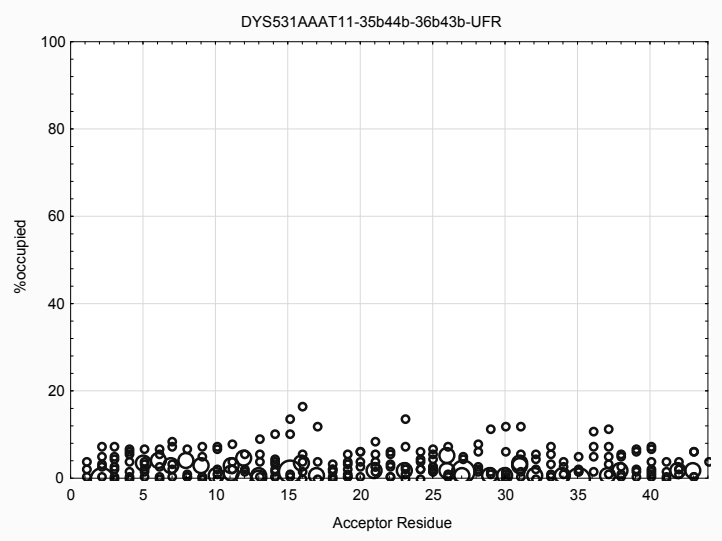
	DYS531AAAT11-27b36b-28b35b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	348	3.18	0.30	85.63	6.72
distance	348	2.82	2.56	3.00	0.08
angle	348	24.20	1.22	58.41	8.99
lifetime	348	5.07	4.00	21.20	1.92
maxocc	348	2.21	1.00	31.00	2.53



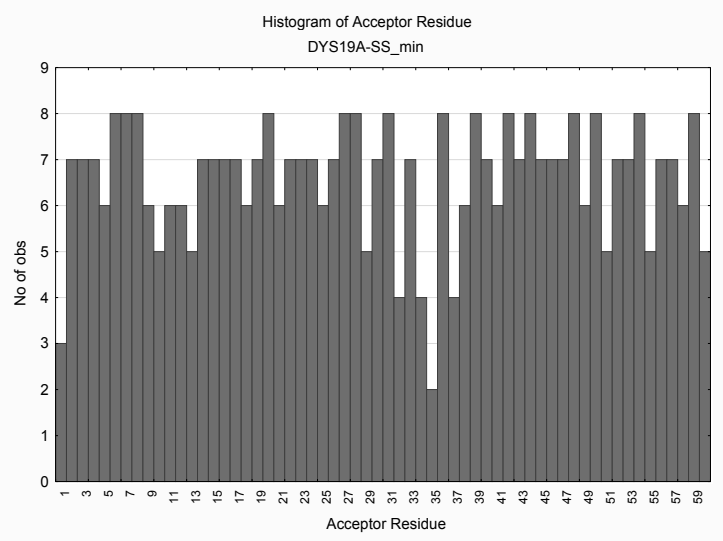
	DYS531AAAT11-31b40b-32b39b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	319	2.44	0.30	20.96	2.57
distance	319	2.82	2.64	2.99	0.08
angle	319	24.79	4.12	57.71	9.24
lifetime	319	5.10	4.00	13.00	1.62
maxocc	319	2.10	1.00	12.00	1.58

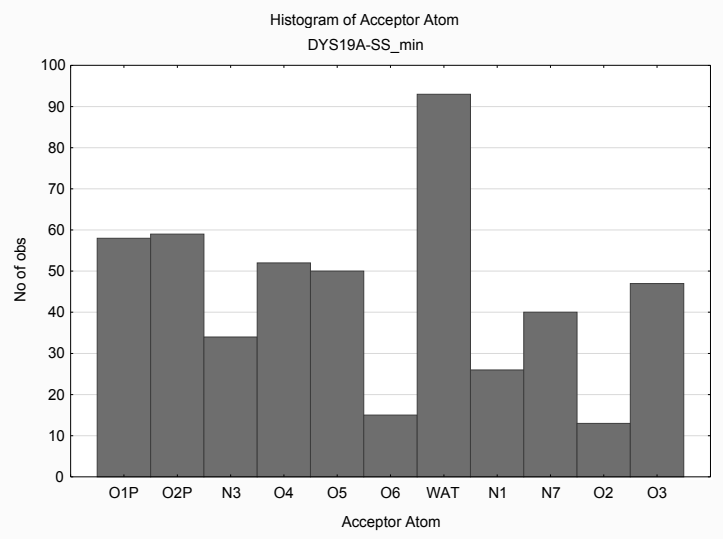


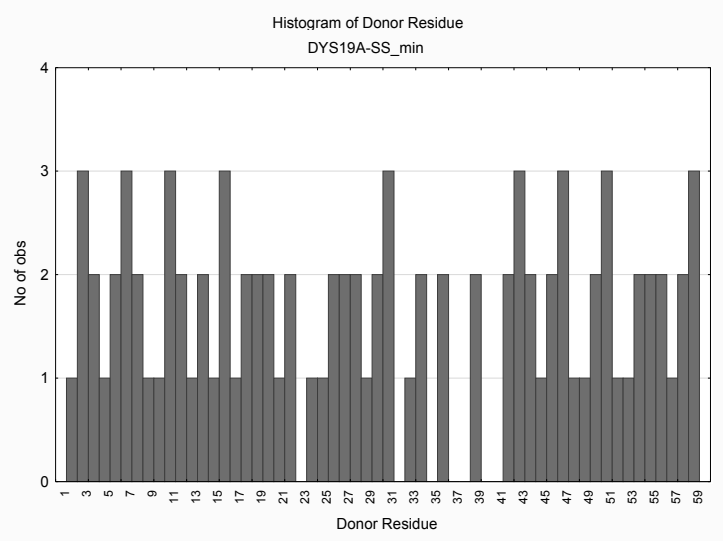
	DYS531AAAT11-35b44b-36b43b-UFR				
Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
%occupied	339	2.90	0.30	16.47	2.78
distance	339	2.82	2.53	3.00	0.08
angle	339	24.91	5.01	59.05	8.76
lifetime	339	5.17	4.00	20.80	2.00
maxocc	339	2.20	1.00	19.00	1.95

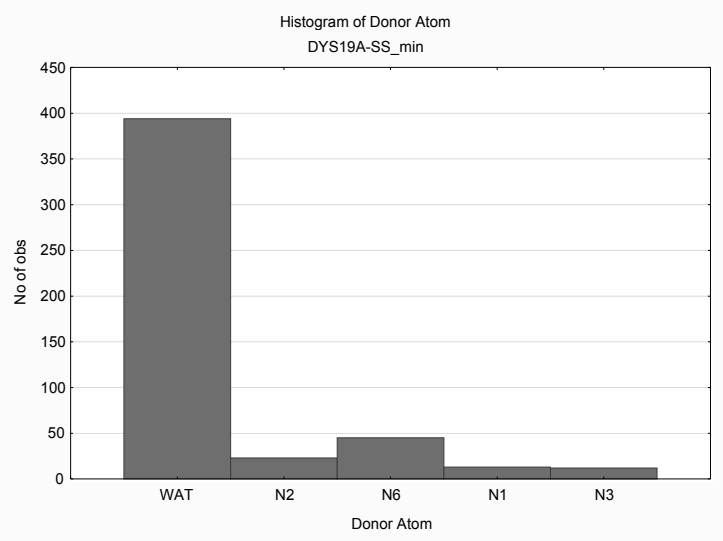


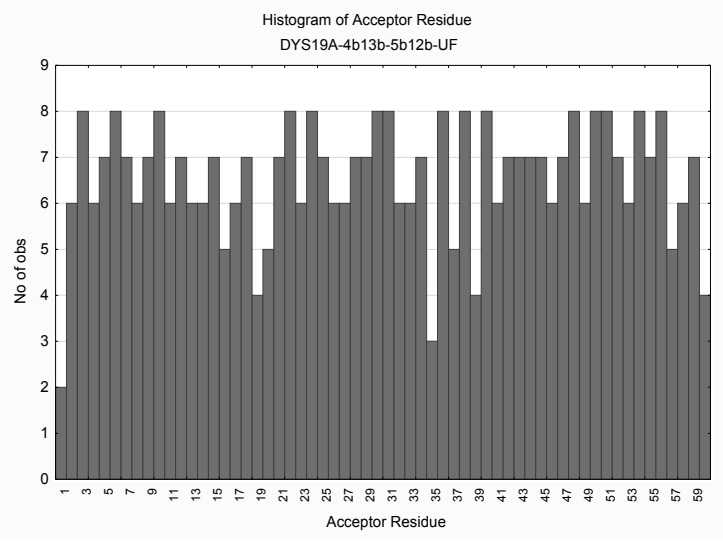
Graphs SI170-SI213: Histograms of water H-bonds (WAT; cut-off 3 Å) near nucleotides (acceptor and donor residue positions) and H-bonds between water and specific atom types (oxygen-phosphate, oxygen and nitrogen atoms), observed for the last 4 ns of molecular dynamics of DYS19A tested models.

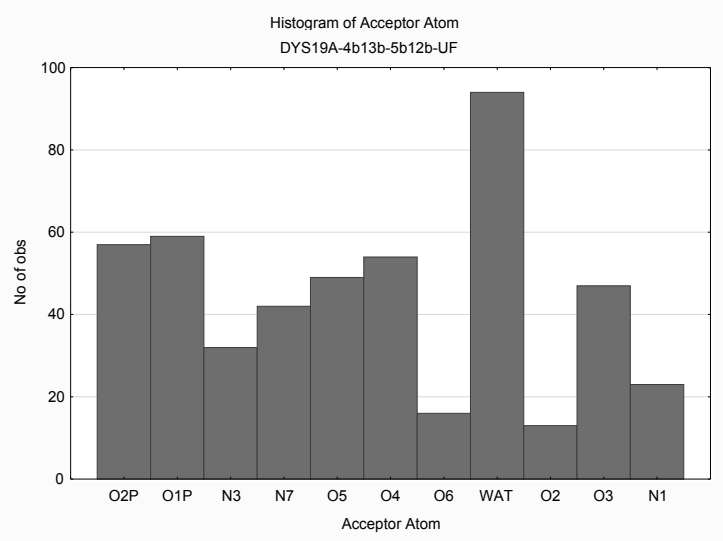


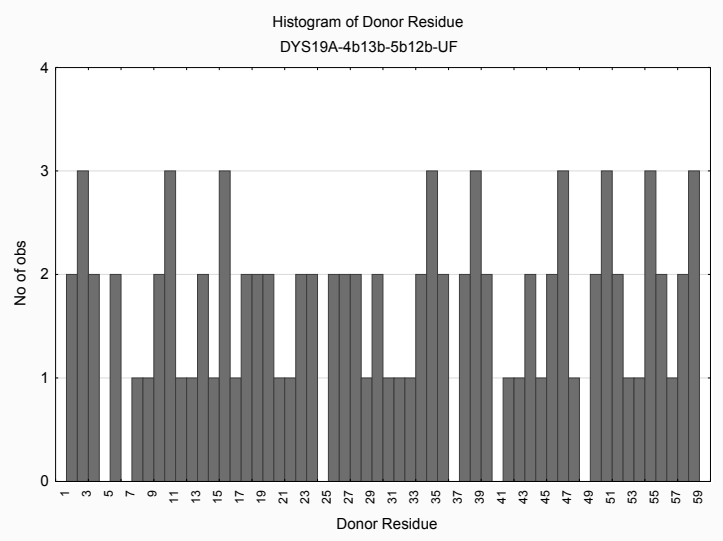


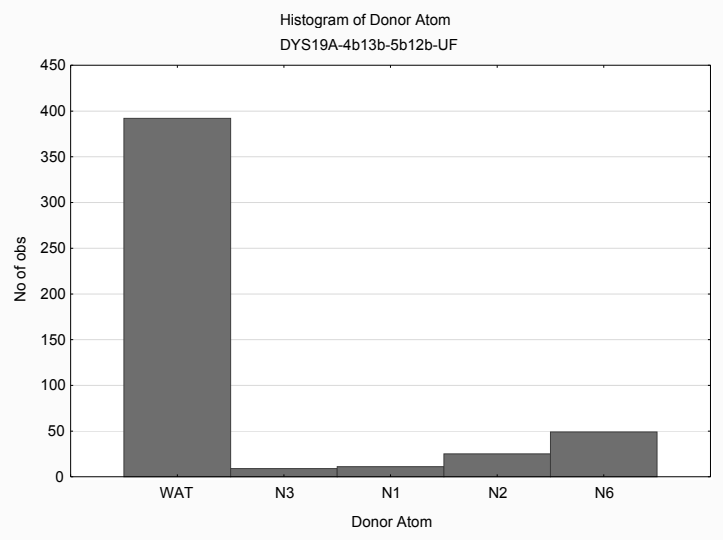

Histogram of Donor Residue
DYS19A-SS_min
No of obs
Donor Residue

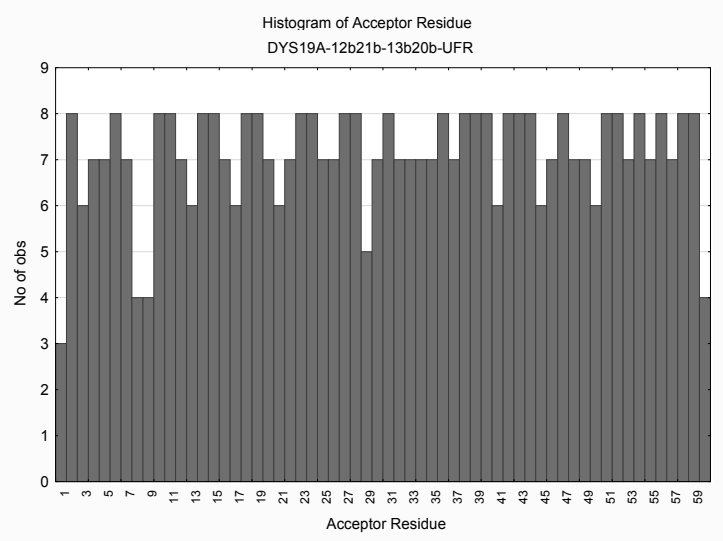

Histogram of Donor Atom
DYS19A-SS_min
No of obs
WAT
N2
N6
N1
N3
Donor Atom

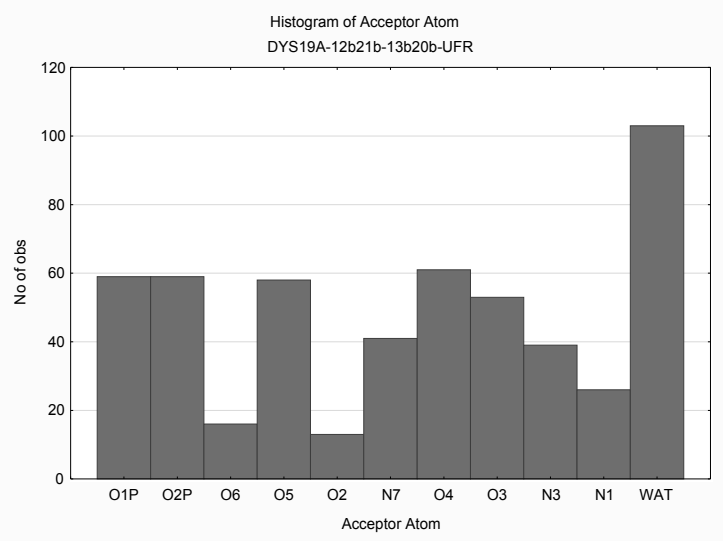

Histogram of Acceptor Residue
DYS19A-4b13b-5b12b-UF
No of obs
Acceptor Residue

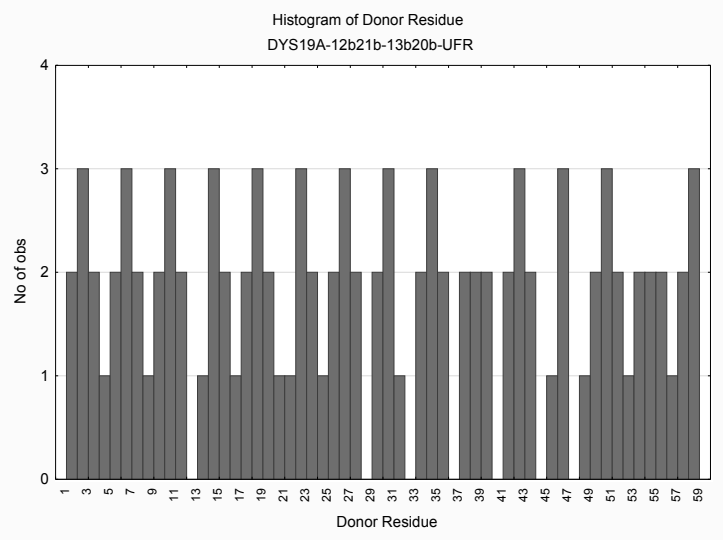

Histogram of Acceptor Atom
DYS19A-4b13b-5b12b-UF
No of obs
0
20
40
60
80
100
O2P
O1P
N3
N7
O5
O4
O6
WAT
O2
O3
N1
Acceptor Atom

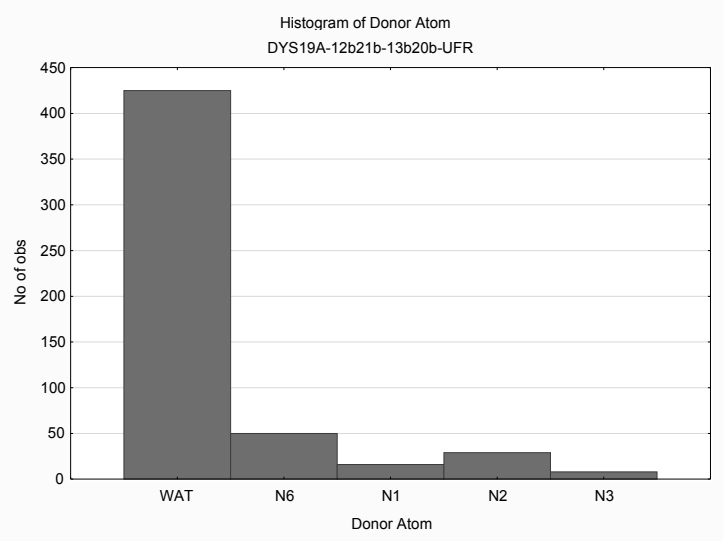

Histogram of Donor Residue
DYS19A-4b13b-5b12b-UF
No of obs
0
1
2
3
4
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
51
53
55
57
59
Donor Residue

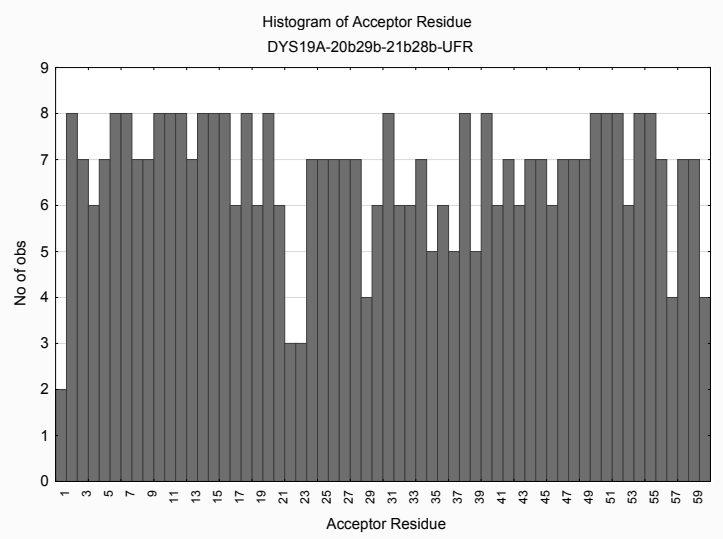

Histogram of Donor Atom
DYS19A-4b13b-5b12b-UF
No of obs
0
50
100
150
200
250
300
350
400
450
WAT
N3
N1
N2
N6
Donor Atom

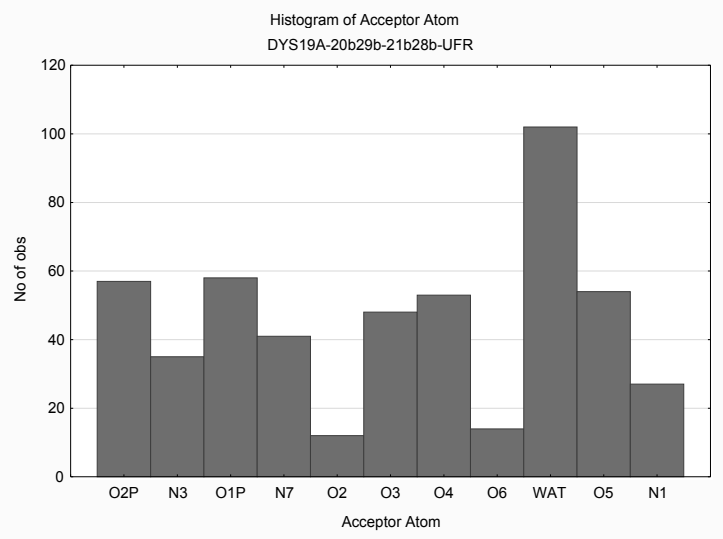

Histogram of Acceptor Residue
DYS19A-12b21b-13b20b-UFR
No of obs
Acceptor Residue

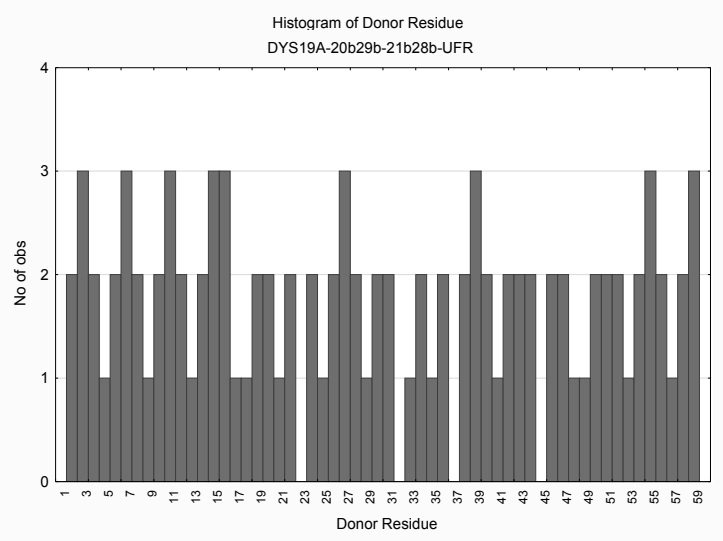

Histogram of Acceptor Atom
DYS19A-12b21b-13b20b-UFR
No of obs
O1P
O2P
O6
O5
O2
N7
O4
O3
N3
N1
WAT
Acceptor Atom

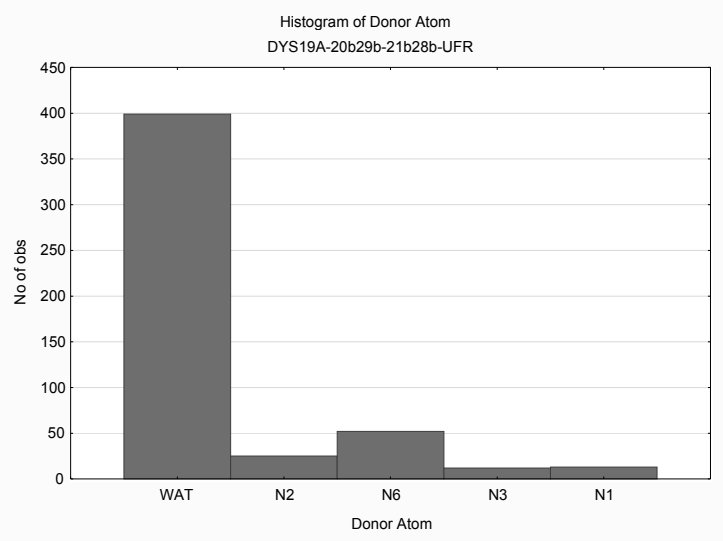

Histogram of Donor Residue
DYS19A-12b21b-13b20b-UFR
No of obs
Donor Residue

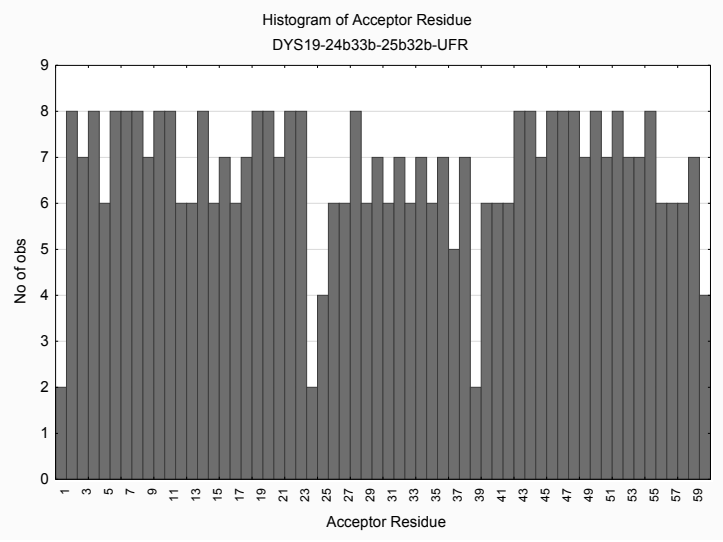

Histogram of Donor Atom
DYS19A-12b21b-13b20b-UFR
No of obs
Donor Atom
WAT
N6
N1
N2
N3

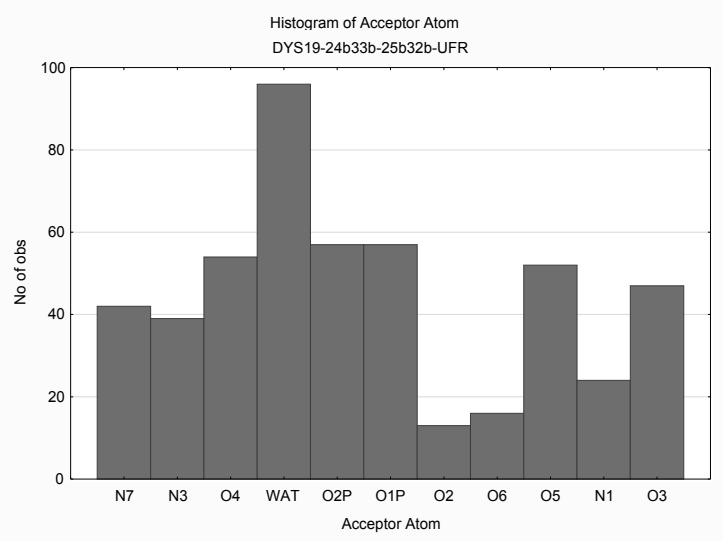

Histogram of Acceptor Residue
DYS19A-20b29b-21b28b-UFR
No of obs
Acceptor Residue

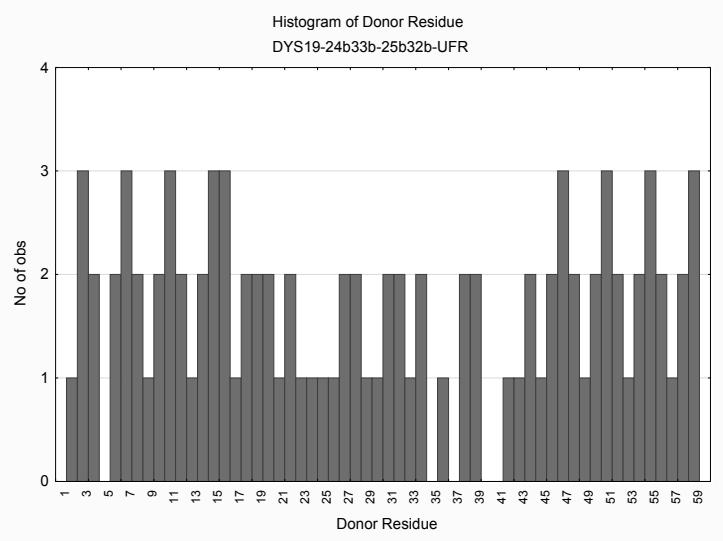

Histogram of Acceptor Atom
DYS19A-20b29b-21b28b-UFR
No of obs
Acceptor Atom
O2P
N3
O1P
N7
O2
O3
O4
O6
WAT
O5
N1

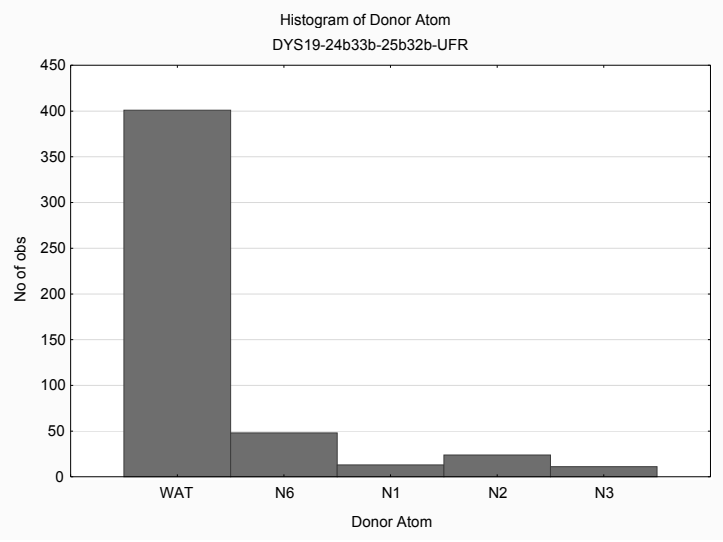

Histogram of Donor Residue
DYS19A-20b29b-21b28b-UFR
No of obs
Donor Residue

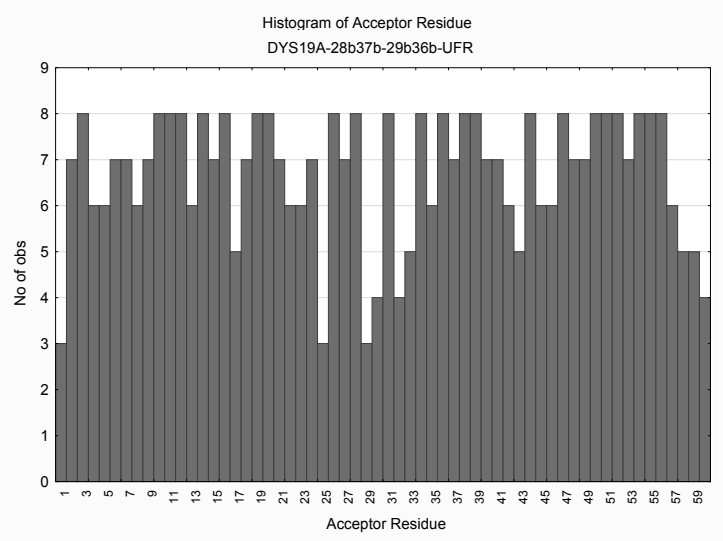

Histogram of Donor Atom
DYS19A-20b29b-21b28b-UFR
No of obs
WAT
N2
N6
N3
N1
Donor Atom

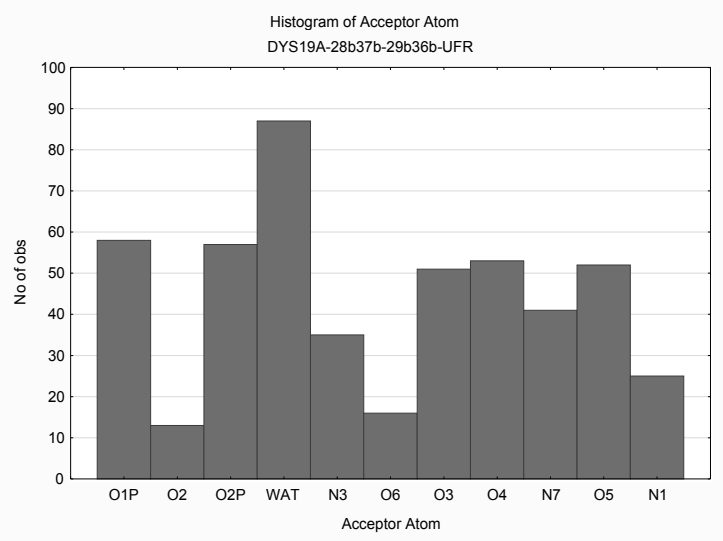

Histogram of Acceptor Residue
DYS19-24b33b-25b32b-UFR
No of obs
Acceptor Residue

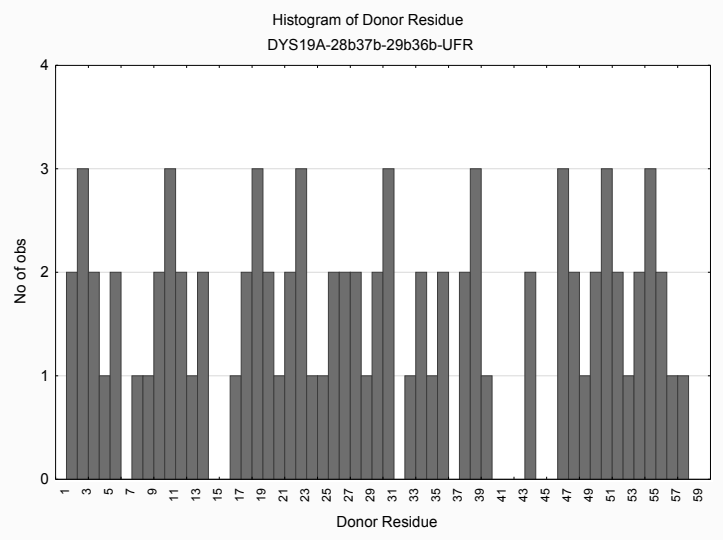

Histogram of Acceptor Atom
DYS19-24b33b-25b32b-UFR
No of obs
0
20
40
60
80
100
N7
N3
O4
WAT
O2P
O1P
O2
O6
O5
N1
O3
Acceptor Atom

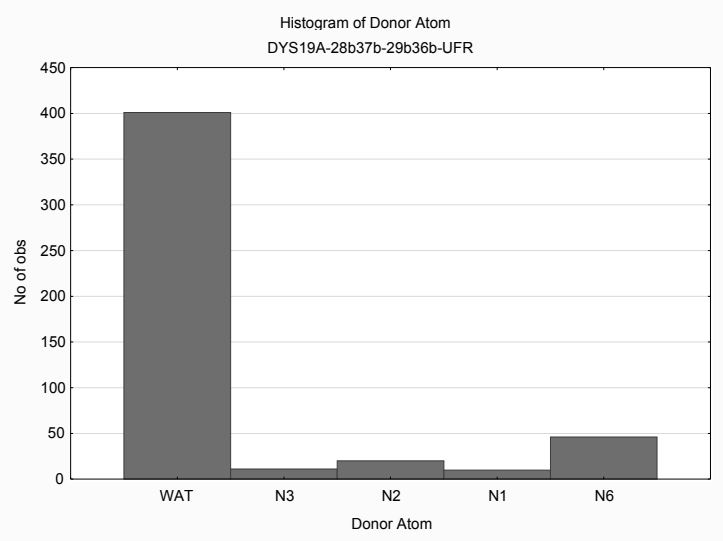

Histogram of Donor Residue
DYS19-24b33b-25b32b-UFR
No of obs
0
1
2
3
4
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
51
53
55
57
59
Donor Residue

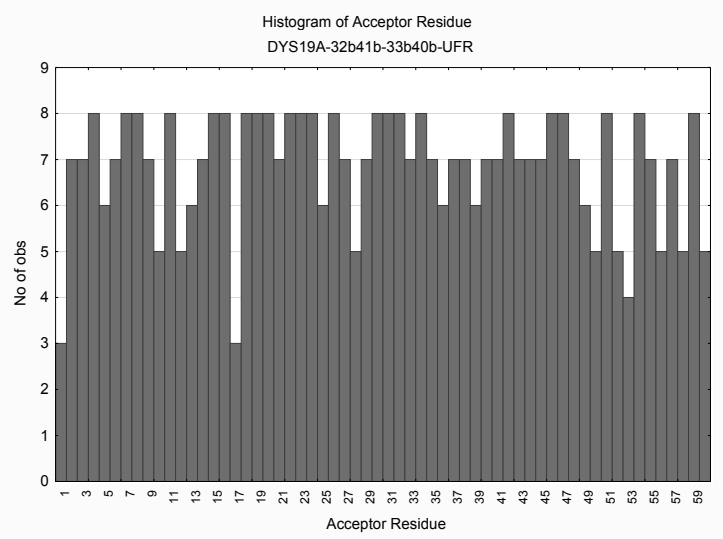

Histogram of Donor Atom
DYS19-24b33b-25b32b-UFR
No of obs
0
50
100
150
200
250
300
350
400
450
WAT
N6
N1
N2
N3
Donor Atom

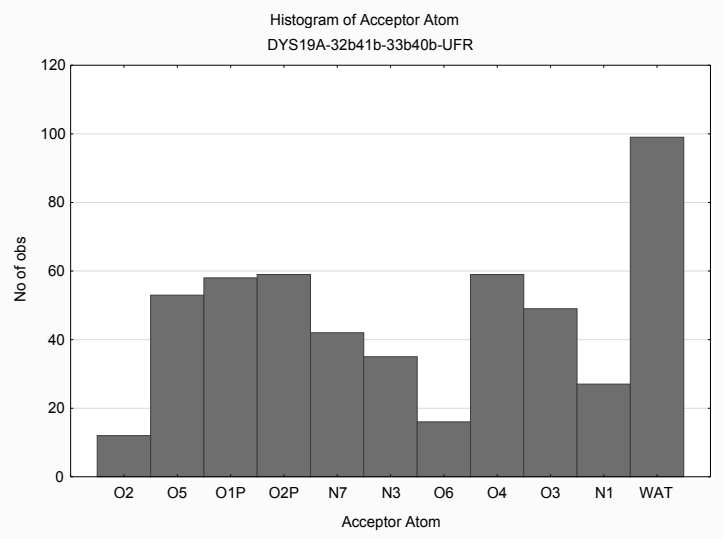

Histogram of Acceptor Residue
DYS19A-28b37b-29b36b-UFR
No of obs
Acceptor Residue

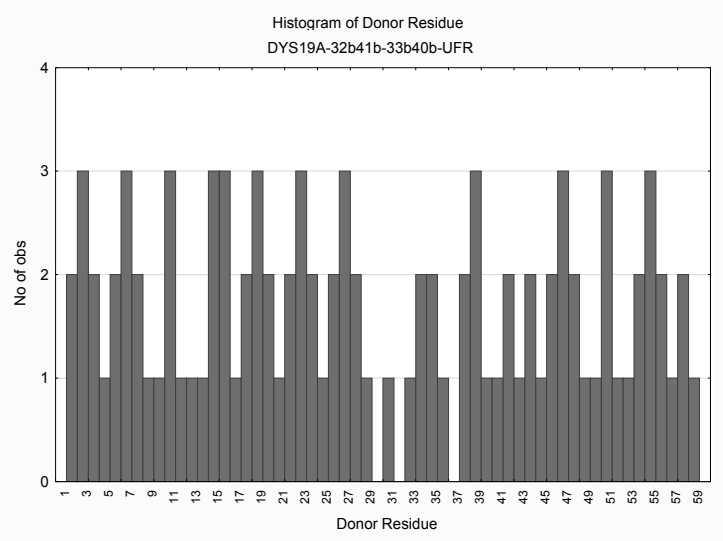

Histogram of Acceptor Atom
DYS19A-28b37b-29b36b-UFR
No of obs
O1P
O2
O2P
WAT
N3
O6
O3
O4
N7
O5
N1
Acceptor Atom

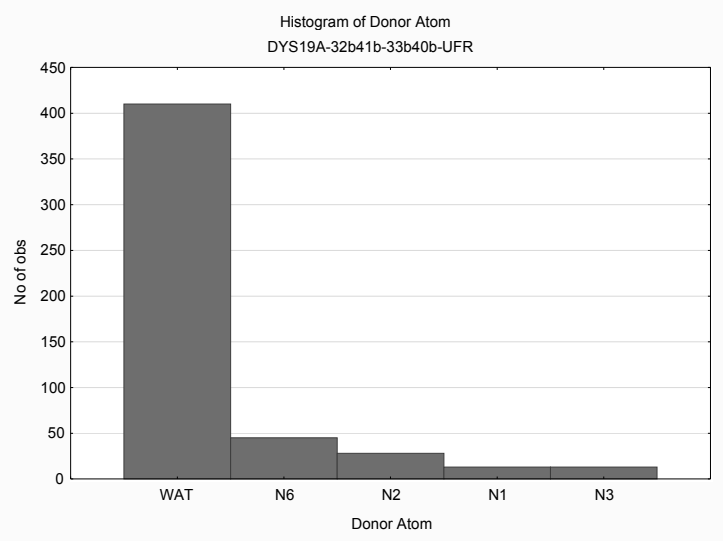

Histogram of Donor Residue
DYS19A-28b37b-29b36b-UFR
No of obs
Donor Residue

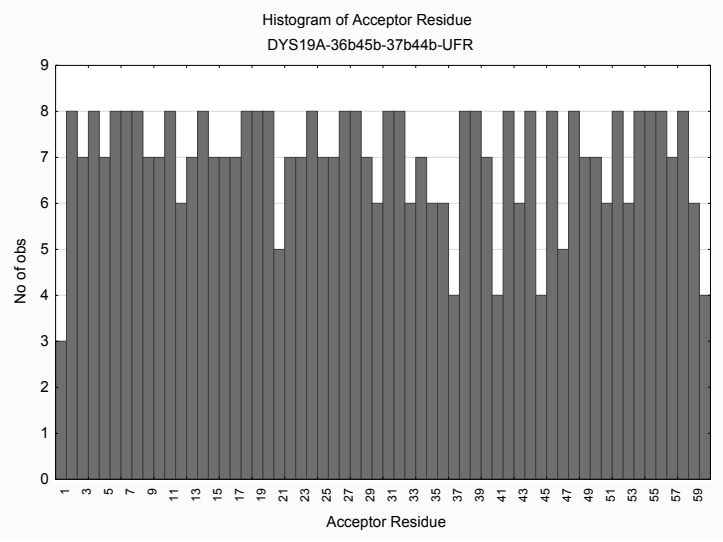

Histogram of Donor Atom
DYS19A-28b37b-29b36b-UFR
No of obs
Donor Atom
WAT
N3
N2
N1
N6

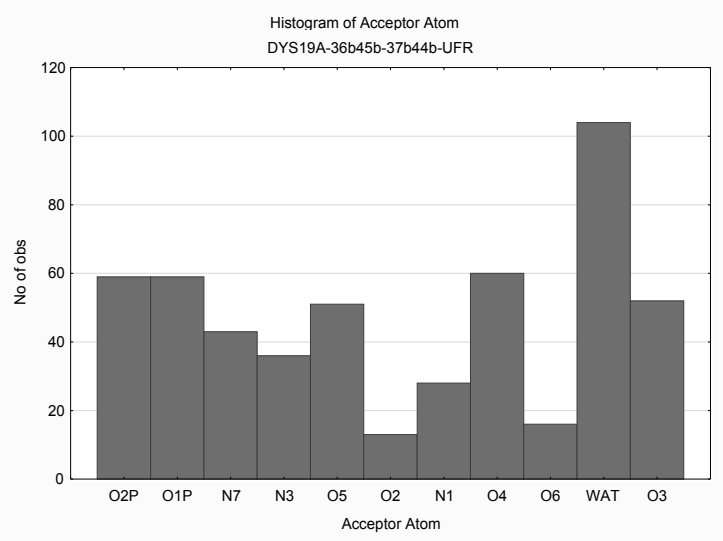

Histogram of Acceptor Residue
DYS19A-32b41b-33b40b-UFR
No of obs
Acceptor Residue

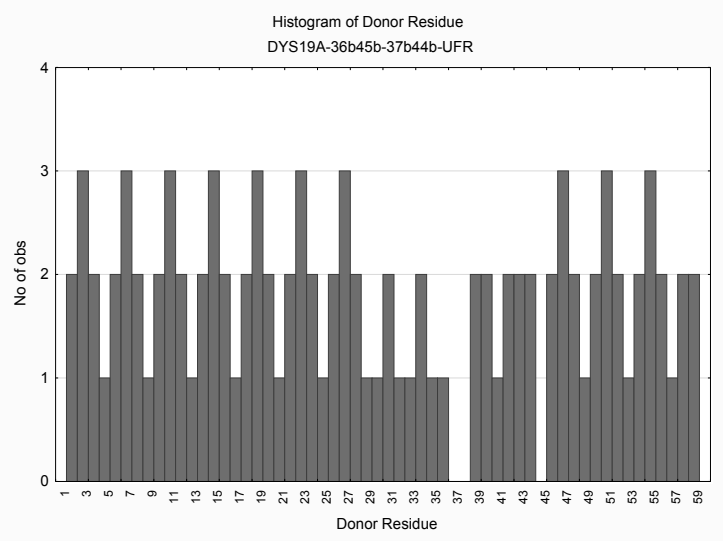

Histogram of Acceptor Atom
DYS19A-32b41b-33b40b-UFR
No of obs
Acceptor Atom
O2
O5
O1P
O2P
N7
N3
O6
O4
O3
N1
WAT

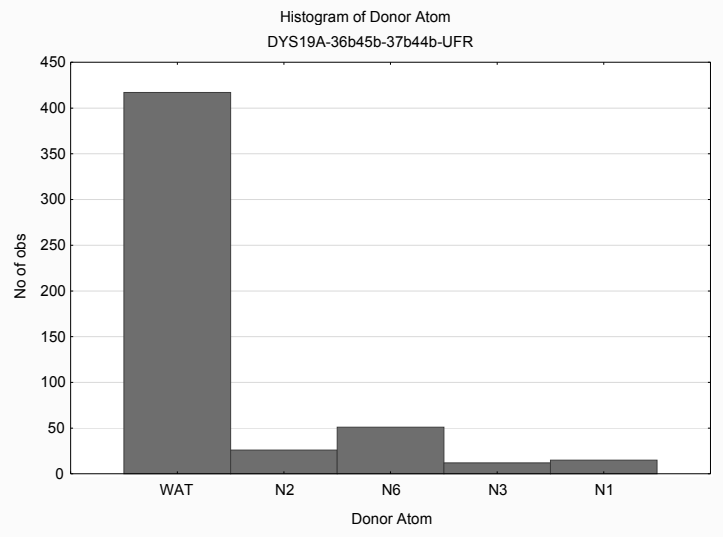

Histogram of Donor Residue
DYS19A-32b41b-33b40b-UFR
No of obs
Donor Residue

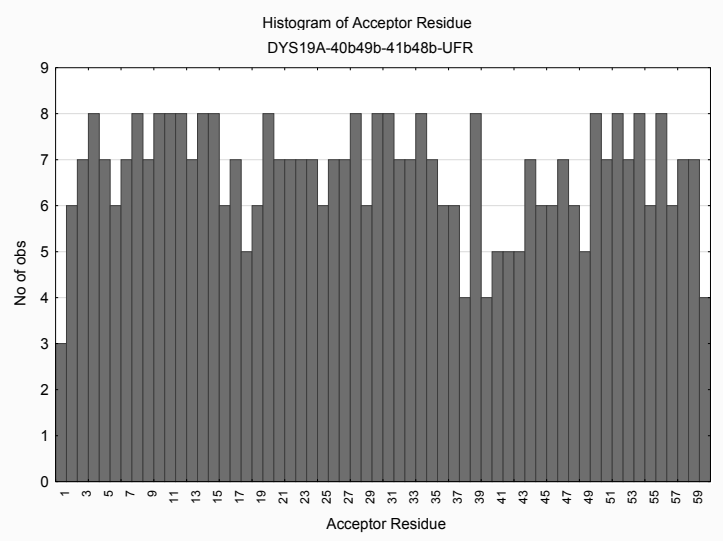

Histogram of Donor Atom
DYS19A-32b41b-33b40b-UFR
No of obs
WAT
N6
N2
N1
N3
Donor Atom

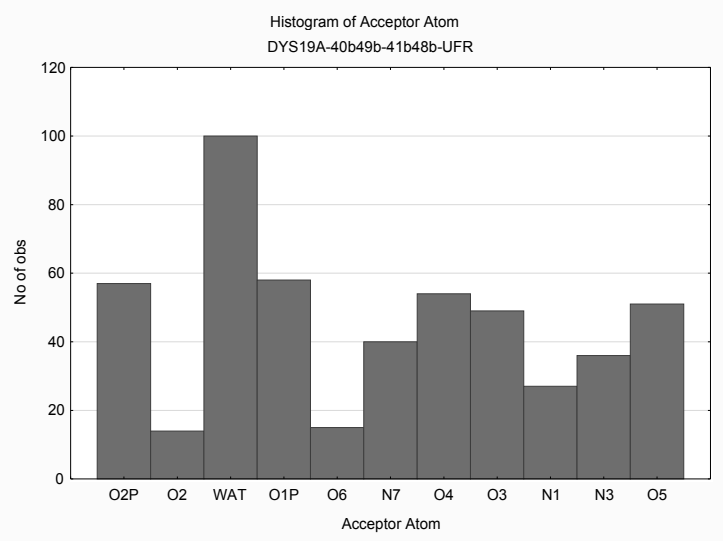

Histogram of Acceptor Residue
DYS19A-36b45b-37b44b-UFR
No of obs
Acceptor Residue

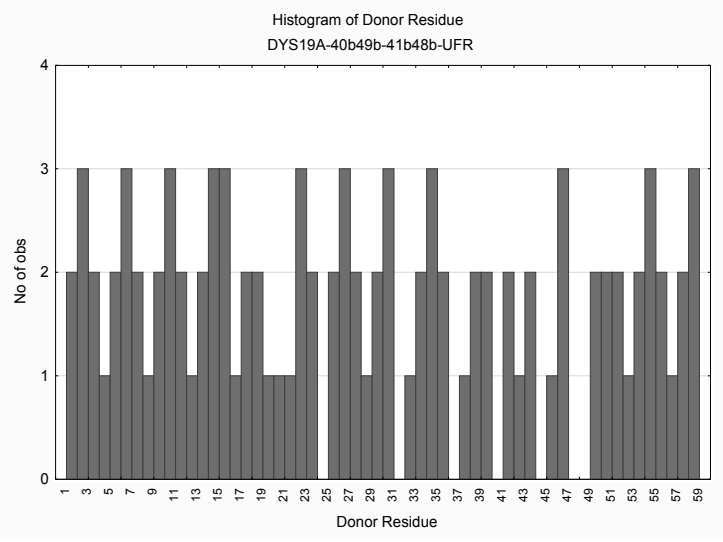

Histogram of Acceptor Atom
DYS19A-36b45b-37b44b-UFR
No of obs
Acceptor Atom
O2P
O1P
N7
N3
O5
O2
N1
O4
O6
WAT
O3

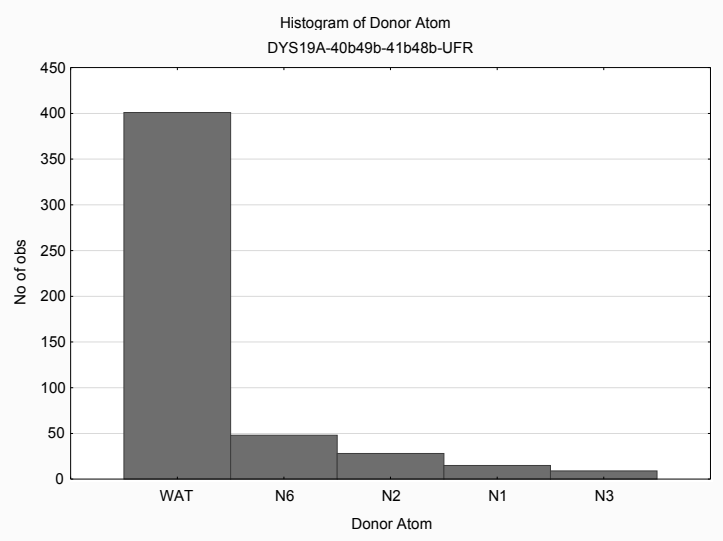

Histogram of Donor Residue
DYS19A-36b45b-37b44b-UFR
No of obs
Donor Residue

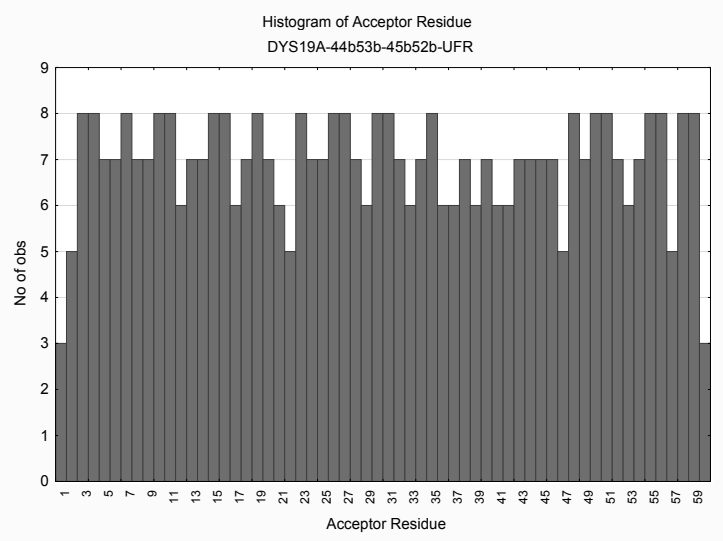

Histogram of Donor Atom
DYS19A-36b45b-37b44b-UFR
No of obs
Donor Atom
WAT
N2
N6
N3
N1

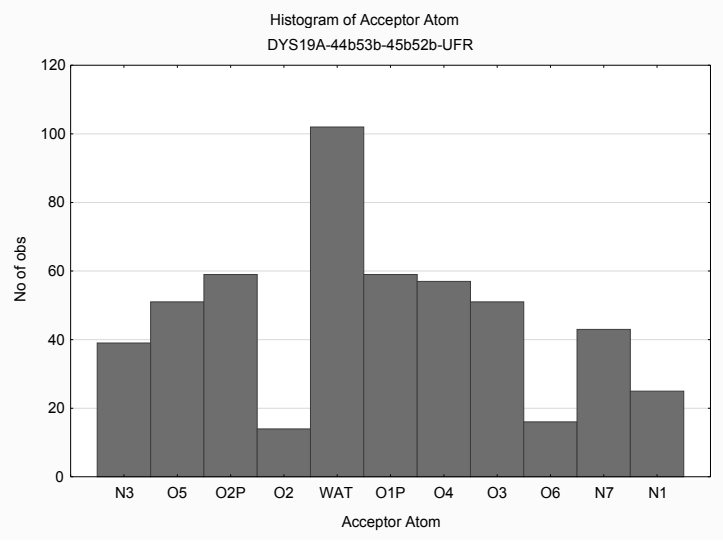

Histogram of Acceptor Residue
DYS19A-40b49b-41b48b-UFR
No of obs
Acceptor Residue

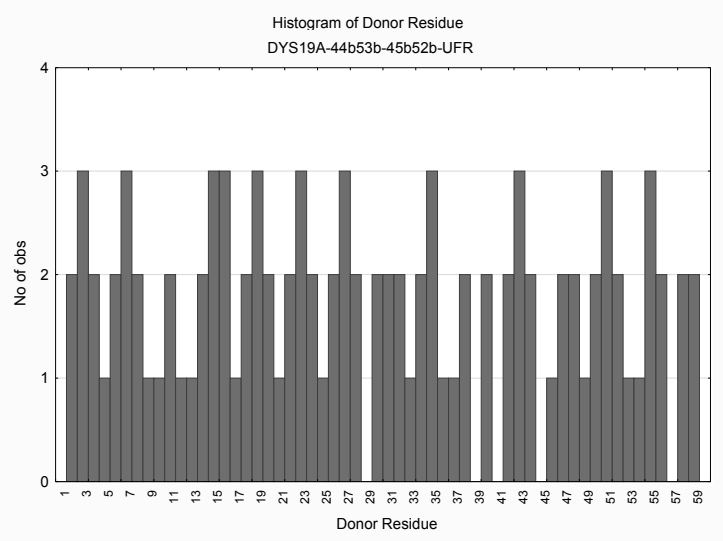

Histogram of Acceptor Atom
DYS19A-40b49b-41b48b-UFR
No of obs
O2P
O2
WAT
O1P
O6
N7
O4
O3
N1
N3
O5
Acceptor Atom

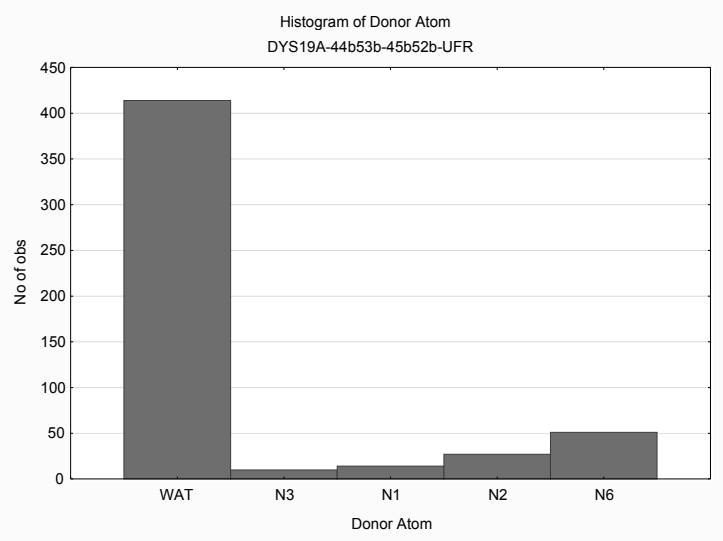

Histogram of Donor Residue
DYS19A-40b49b-41b48b-UFR
No of obs
Donor Residue

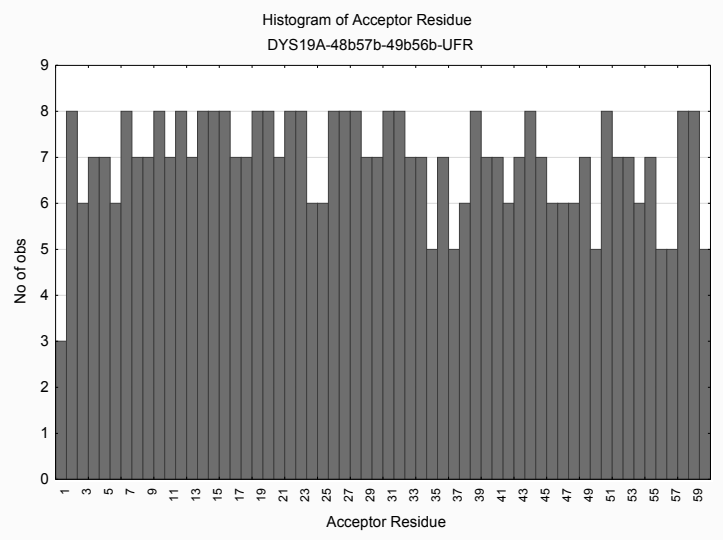

Histogram of Donor Atom
DYS19A-40b49b-41b48b-UFR
No of obs
0
50
100
150
200
250
300
350
400
450
WAT
N6
N2
N1
N3
Donor Atom

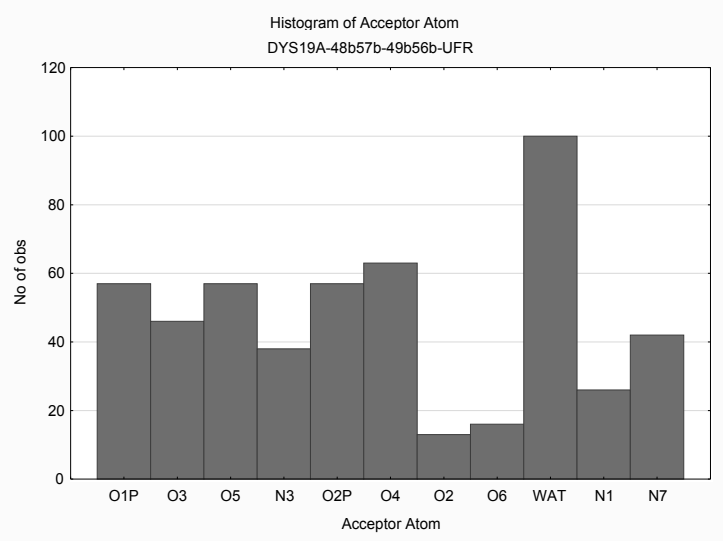

Histogram of Acceptor Residue
DYS19A-44b53b-45b52b-UFR
No of obs
0
1
2
3
4
5
6
7
8
9
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
51
53
55
57
59
Acceptor Residue

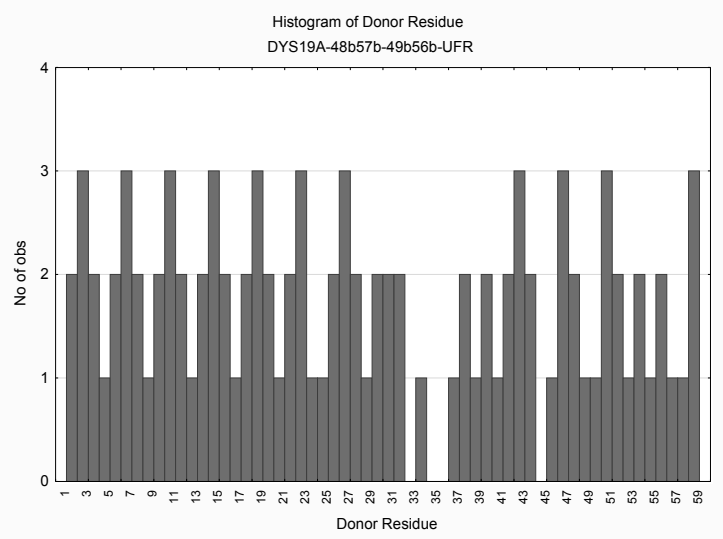

Histogram of Acceptor Atom
DYS19A-44b53b-45b52b-UFR
No of obs
0
20
40
60
80
100
120
N3
O5
O2P
O2
WAT
O1P
O4
O3
O6
N7
N1
Acceptor Atom

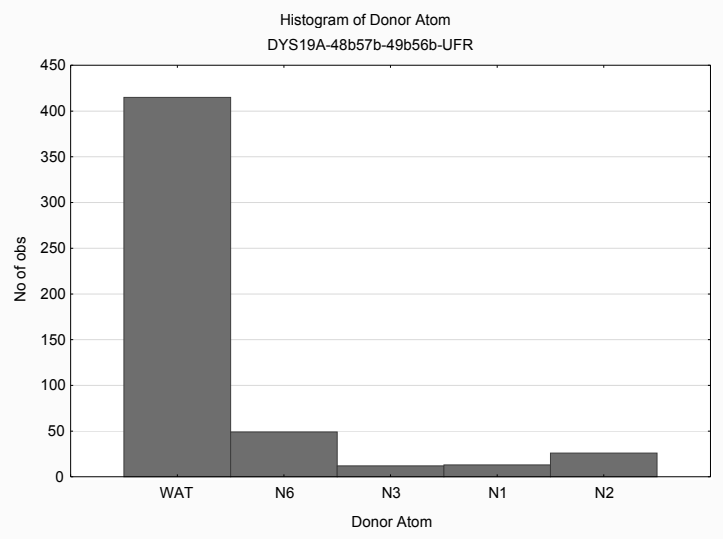

Histogram of Donor Residue
DYS19A-44b53b-45b52b-UFR
No of obs
Donor Residue


Histogram of Donor Atom
DYS19A-44b53b-45b52b-UFR
No of obs
WAT
N3
N1
N2
N6
Donor Atom

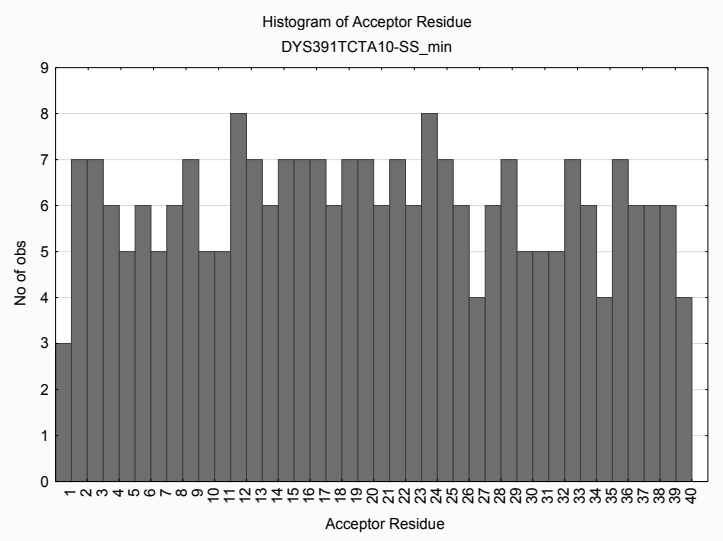

Histogram of Acceptor Residue
DYS19A-48b57b-49b56b-UFR
No of obs
Acceptor Residue

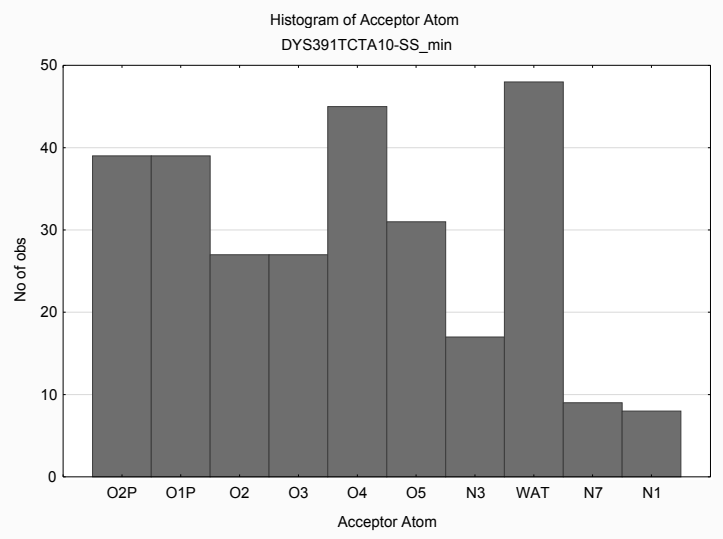

Histogram of Acceptor Atom
DYS19A-48b57b-49b56b-UFR
No of obs
0
20
40
60
80
100
120
O1P
O3
O5
N3
O2P
O4
O2
O6
WAT
N1
N7
Acceptor Atom

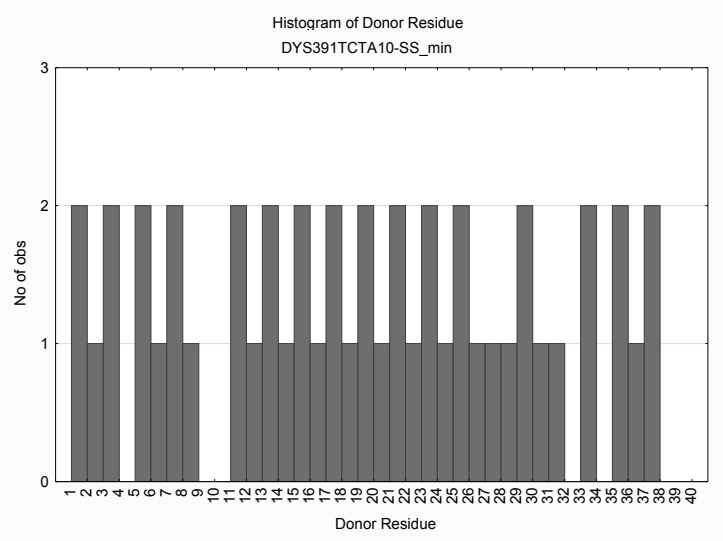

Histogram of Donor Residue
DYS19A-48b57b-49b56b-UFR
No of obs
0
1
2
3
4
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
51
53
55
57
59
Donor Residue

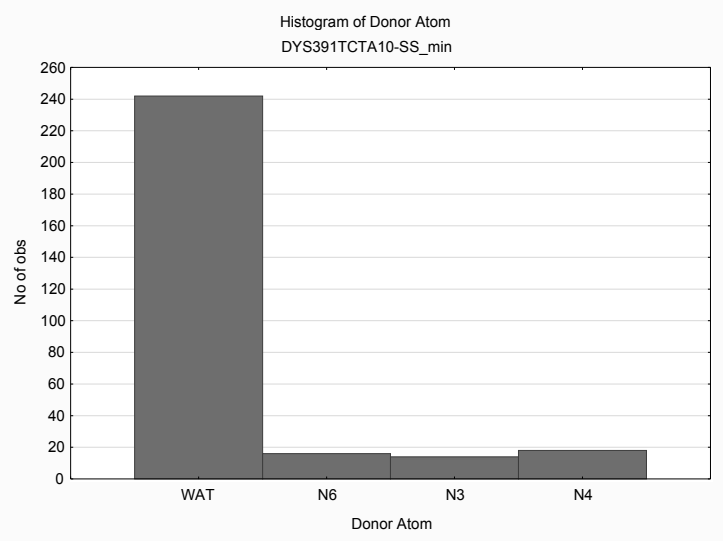

Histogram of Donor Atom
DYS19A-48b57b-49b56b-UFR
No of obs
0
50
100
150
200
250
300
350
400
450
WAT
N6
N3
N1
N2
Donor Atom

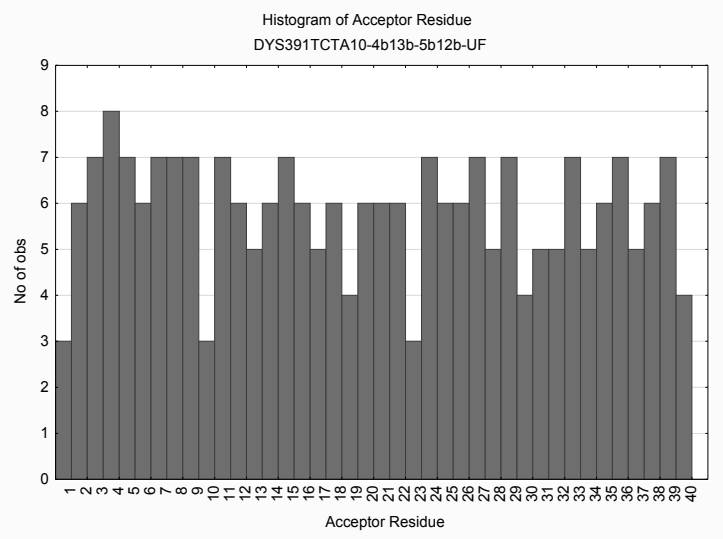
Graphs SI214-SI241: Histograms of water H-bonds (WAT; cut-off 3 Å) near nucleotides (acceptor and donor residue positions) and H-bonds between water and specific atom types (oxygen-phosphate, oxygen and nitrogen atoms), observed for the last 4 ns of molecular dynamics of DYS391 tested models.

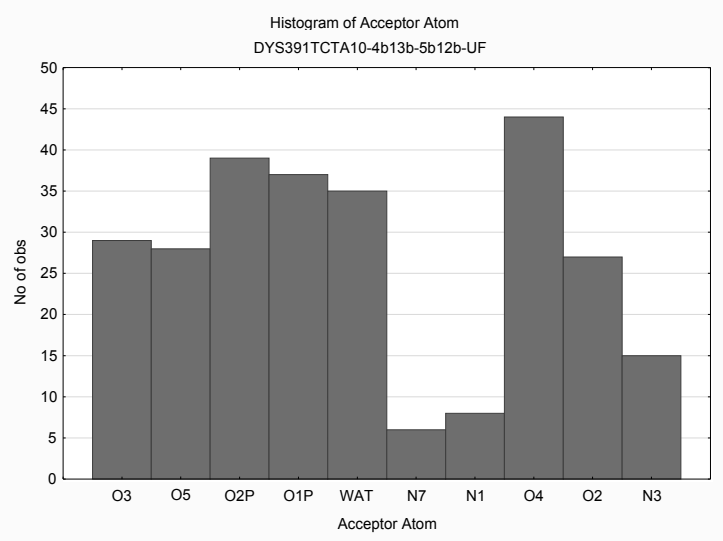


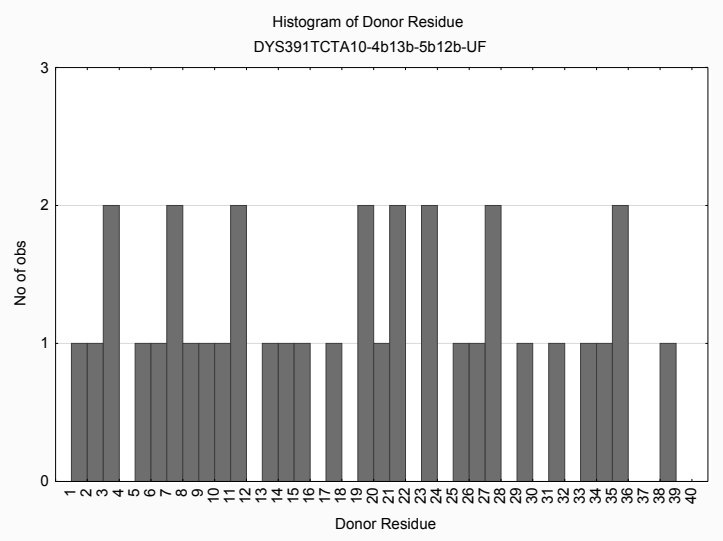


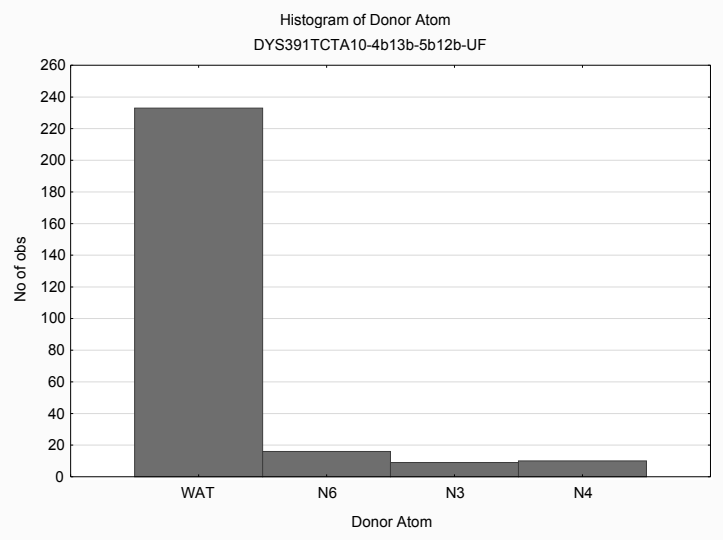

Histogram of Donor Residue
DYS391TCTA10-SS_min
No of obs
Donor Residue

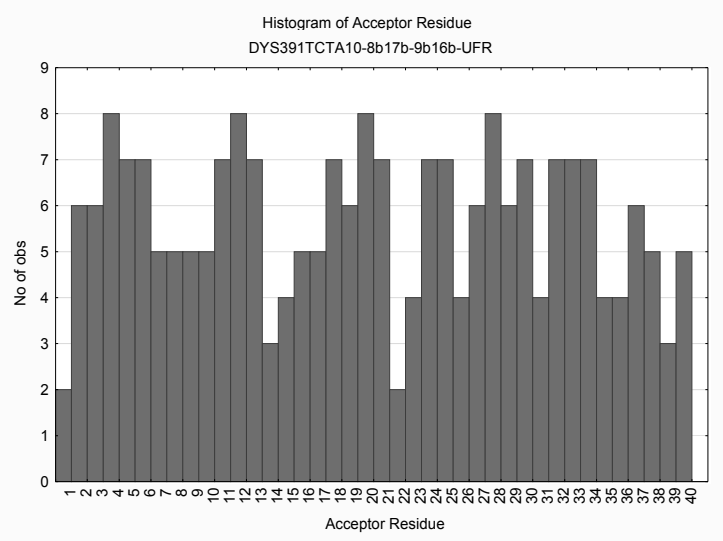

Histogram of Donor Atom
DYS391TCTA10-SS_min
No of obs
WAT
N6
N3
N4
Donor Atom

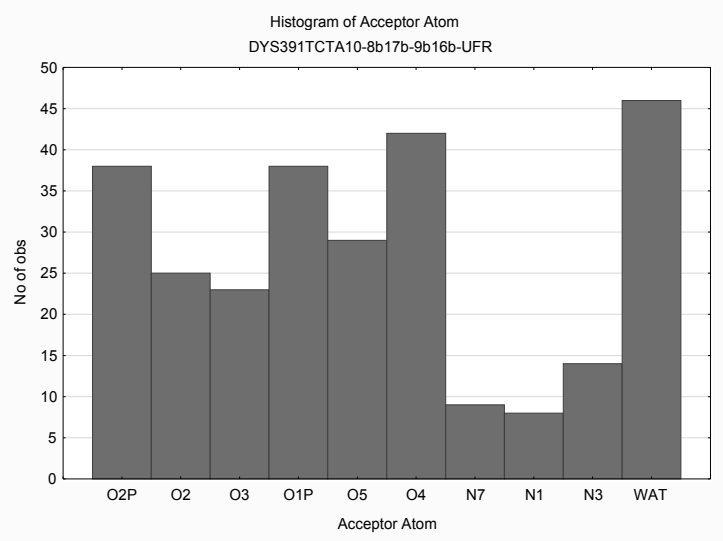

Histogram of Acceptor Residue
DYS391TCTA10-4b13b-5b12b-UF
No of obs
Acceptor Residue

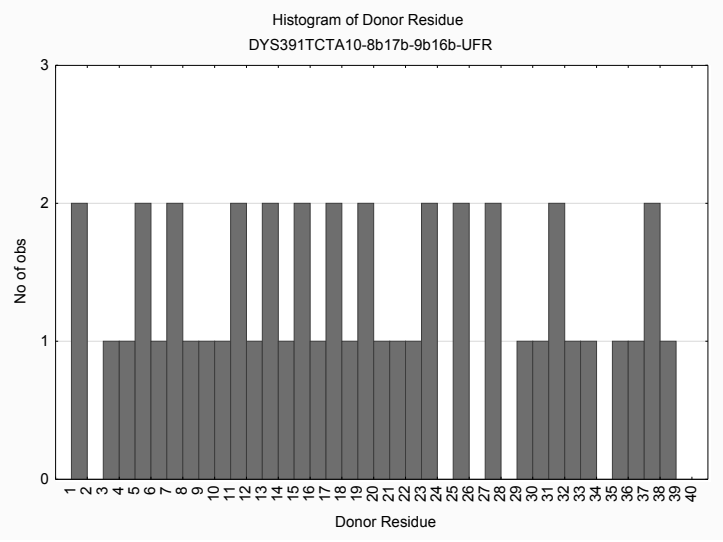

Histogram of Acceptor Atom
DYS391TCTA10-4b13b-5b12b-UF
No of obs
0
5
10
15
20
25
30
35
40
45
50
O3
O5
O2P
O1P
WAT
N7
N1
O4
O2
N3
Acceptor Atom

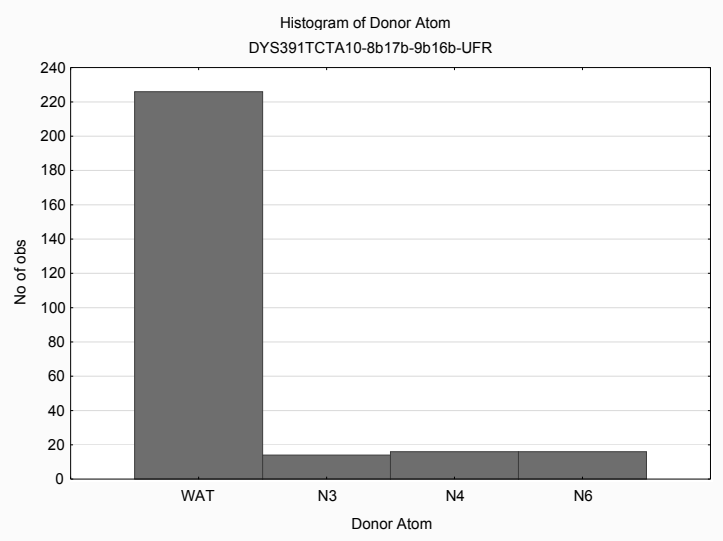

Histogram of Donor Residue
DYS391TCTA10-4b13b-5b12b-UF
No of obs
0
1
2
3
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
Donor Residue

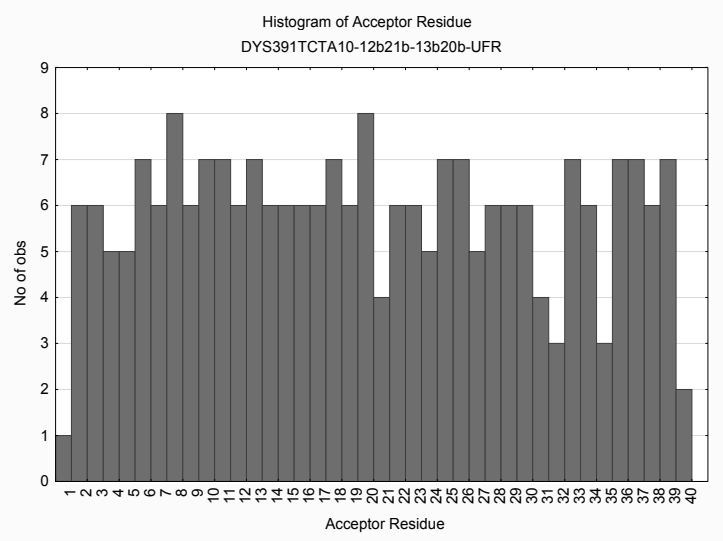

Histogram of Donor Atom
DYS391TCTA10-4b13b-5b12b-UF
No of obs
0
20
40
60
80
100
120
140
160
180
200
220
240
260
WAT
N6
N3
N4
Donor Atom

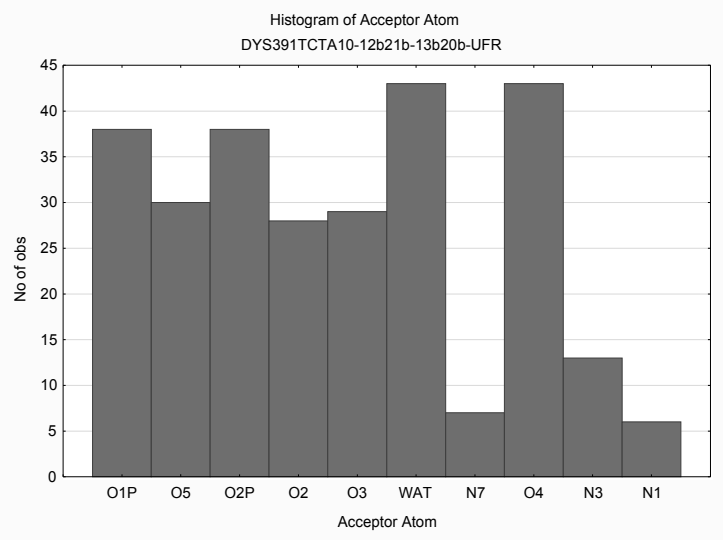

Histogram of Acceptor Residue
DYS391TCTA10-8b17b-9b16b-UFR
No of obs
Acceptor Residue

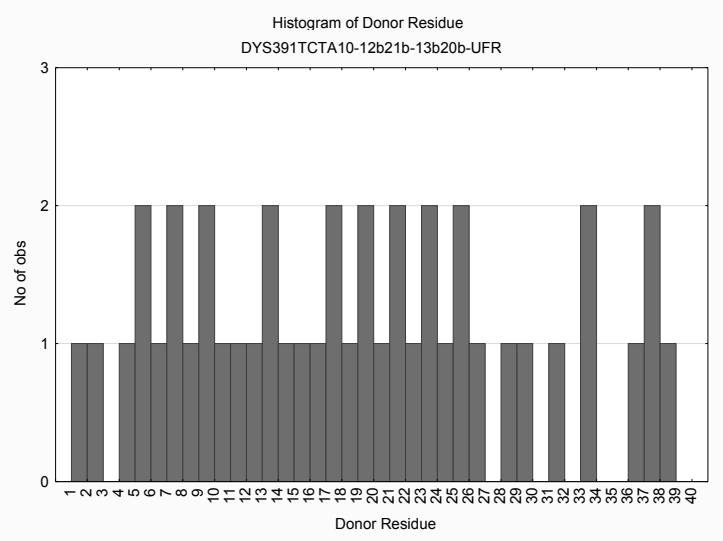

Histogram of Acceptor Atom
DYS391TCTA10-8b17b-9b16b-UFR
No of obs
O2P
O2
O3
O1P
O5
O4
N7
N1
N3
WAT
Acceptor Atom

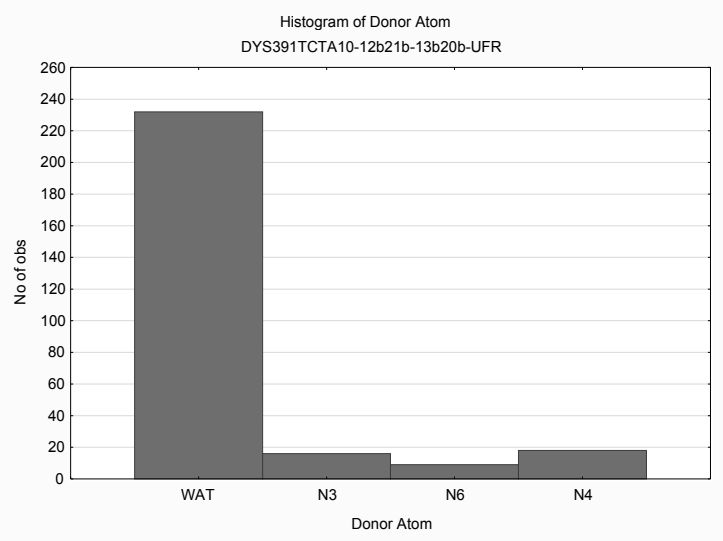

Histogram of Donor Residue
DYS391TCTA10-8b17b-9b16b-UFR
No of obs
Donor Residue

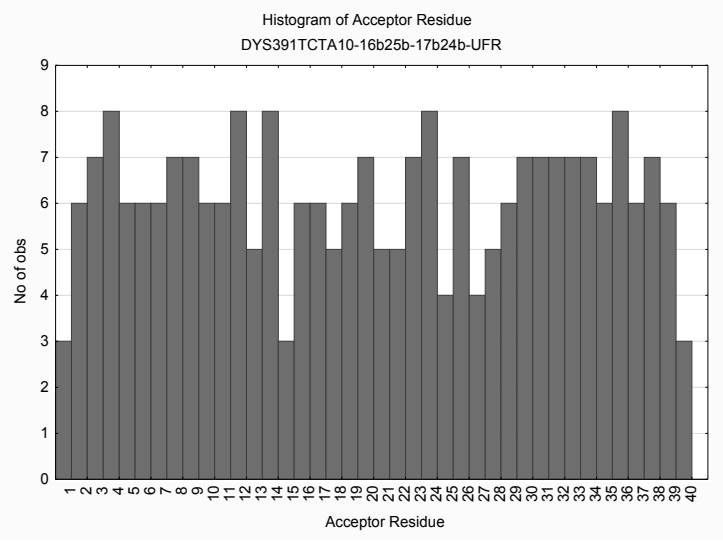

Histogram of Donor Atom
DYS391TCTA10-8b17b-9b16b-UFR
No of obs
0
20
40
60
80
100
120
140
160
180
200
220
240
WAT
N3
N4
N6
Donor Atom

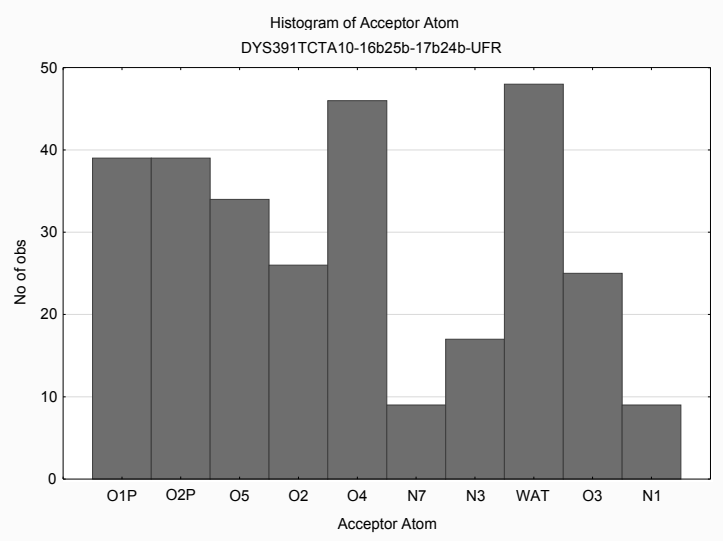

Histogram of Acceptor Residue
DYS391TCTA10-12b21b-13b20b-UFR
No of obs
0
1
2
3
4
5
6
7
8
9
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
Acceptor Residue

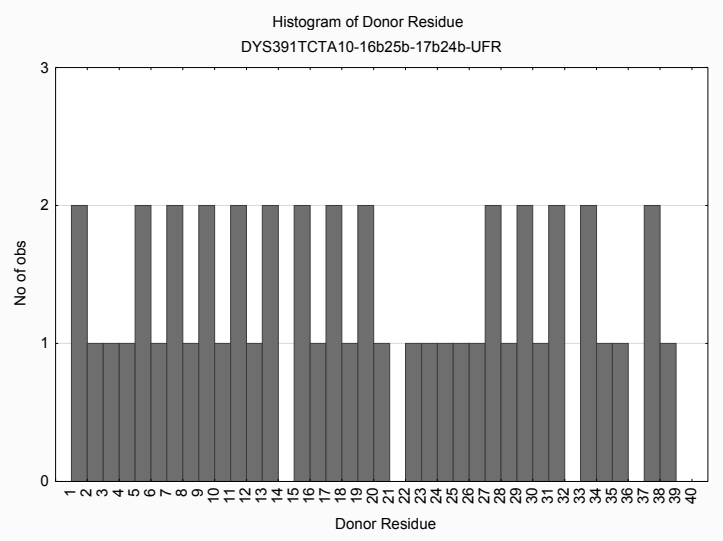

Histogram of Acceptor Atom
DYS391TCTA10-12b21b-13b20b-UFR
No of obs
0
5
10
15
20
25
30
35
40
45
O1P
O5
O2P
O2
O3
WAT
N7
O4
N3
N1
Acceptor Atom

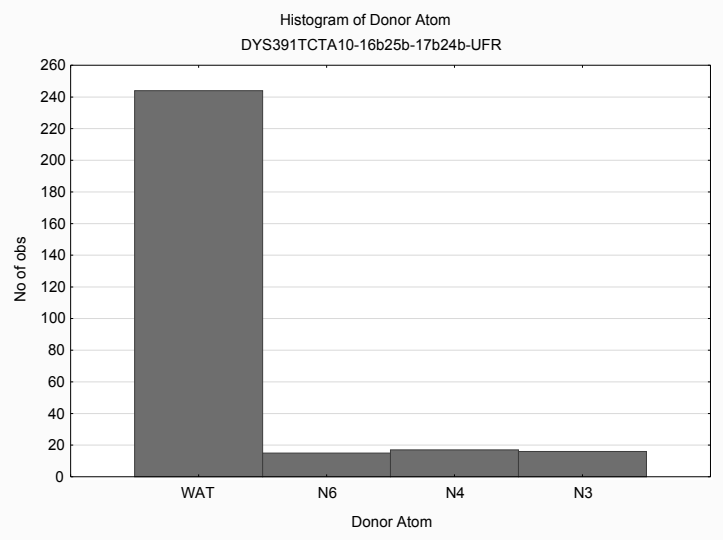

Histogram of Donor Residue
DYS391TCTA10-12b21b-13b20b-UFR
No of obs
Donor Residue

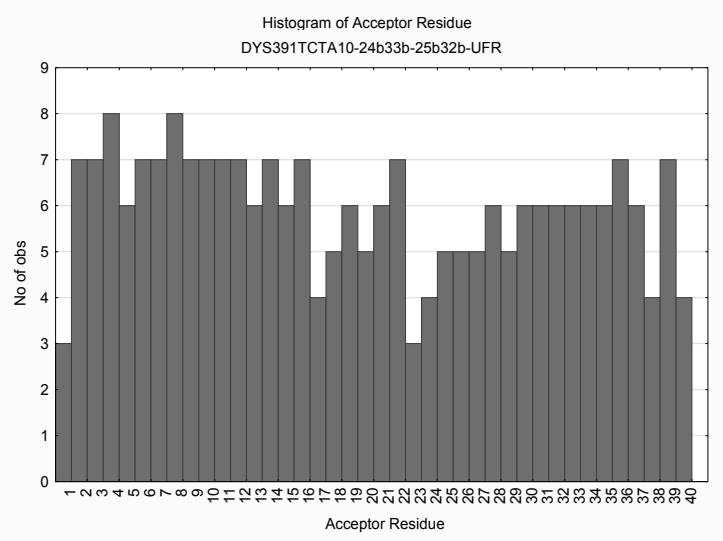

Histogram of Donor Atom
DYS391TCTA10-12b21b-13b20b-UFR
No of obs
WAT
N3
N6
N4
Donor Atom

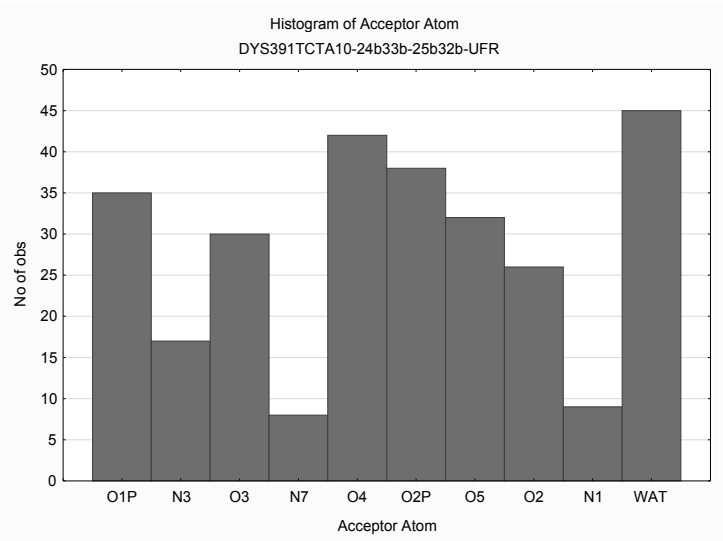

Histogram of Acceptor Residue
DYS391TCTA10-16b25b-17b24b-UFR
No of obs
Acceptor Residue

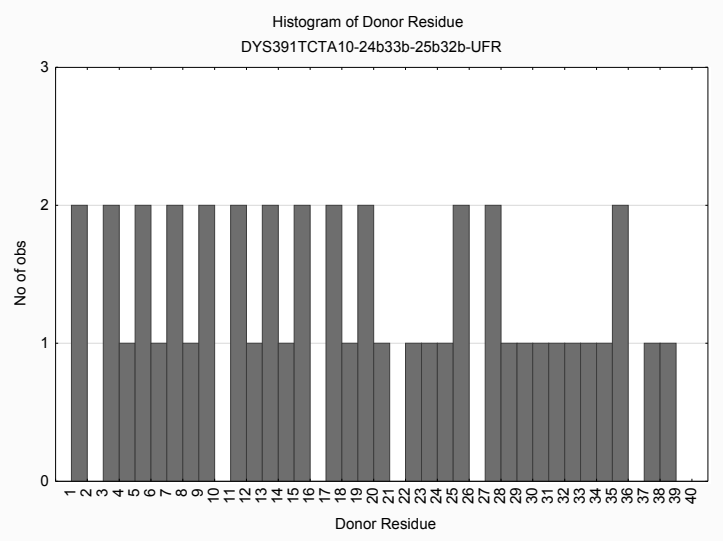

Histogram of Acceptor Atom
DYS391TCTA10-16b25b-17b24b-UFR
No of obs
Acceptor Atom
O1P
O2P
O5
O2
O4
N7
N3
WAT
O3
N1

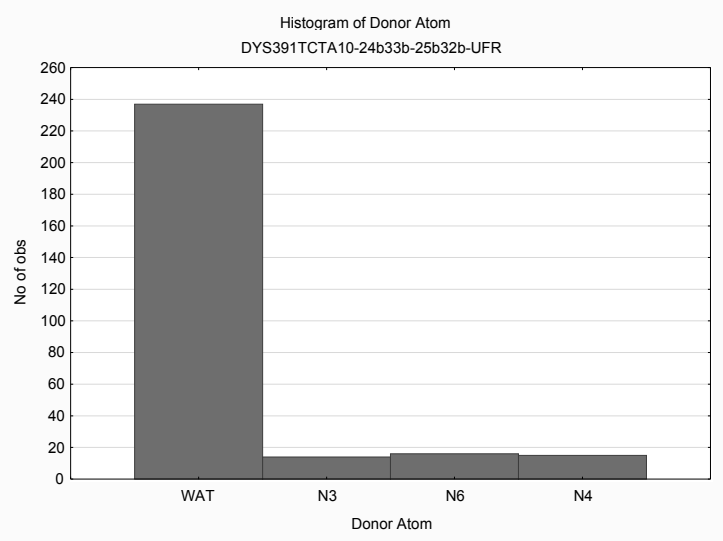

Histogram of Donor Residue
DYS391TCTA10-16b25b-17b24b-UFR
No of obs
Donor Residue

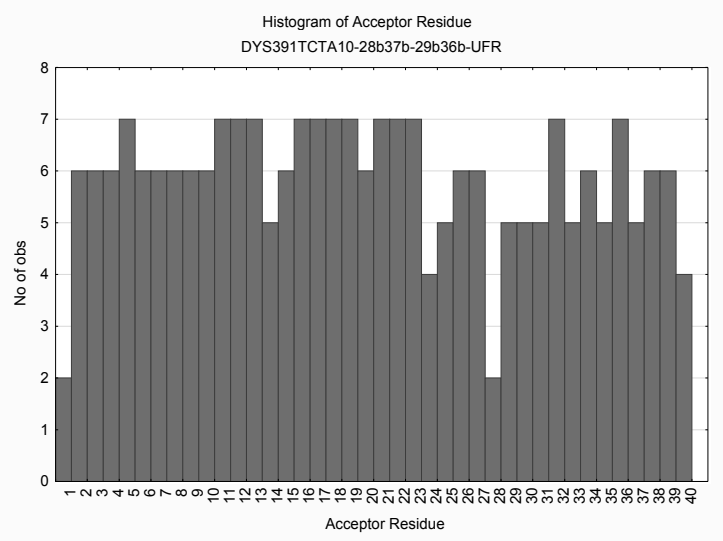

Histogram of Donor Atom
DYS391TCTA10-16b25b-17b24b-UFR
No of obs
Donor Atom
WAT
N6
N4
N3

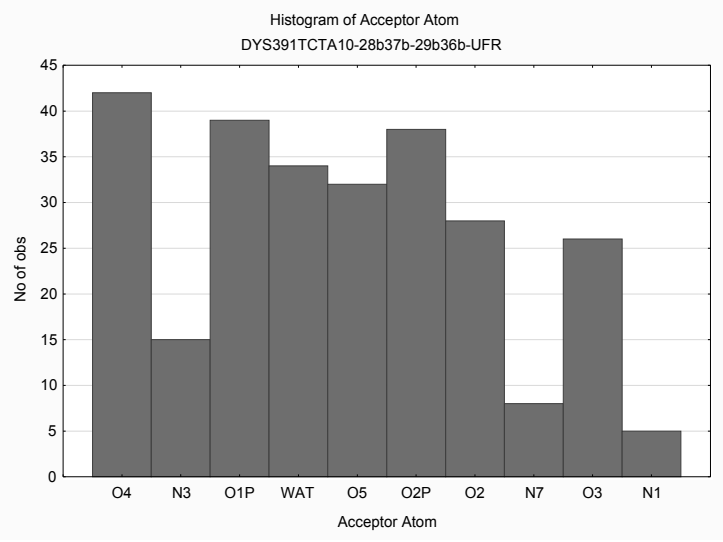

Histogram of Acceptor Residue
DYS391TCTA10-24b33b-25b32b-UFR
No of obs
Acceptor Residue

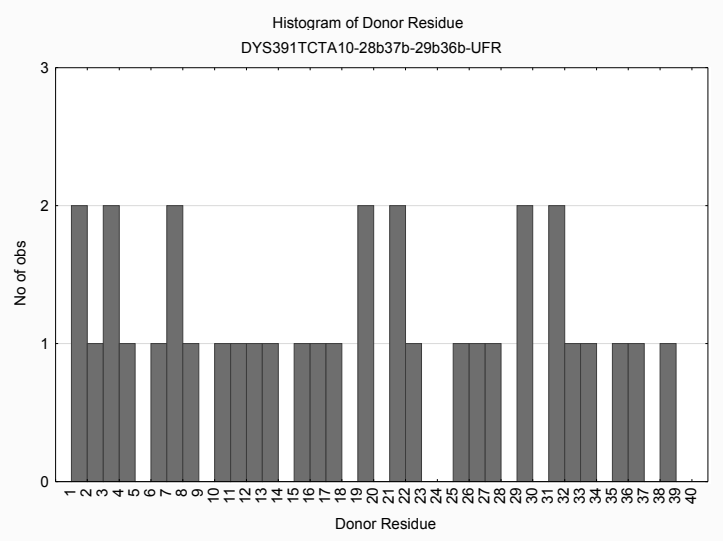

Histogram of Acceptor Atom
DYS391TCTA10-24b33b-25b32b-UFR
No of obs
O1P
N3
O3
N7
O4
O2P
O5
O2
N1
WAT
Acceptor Atom

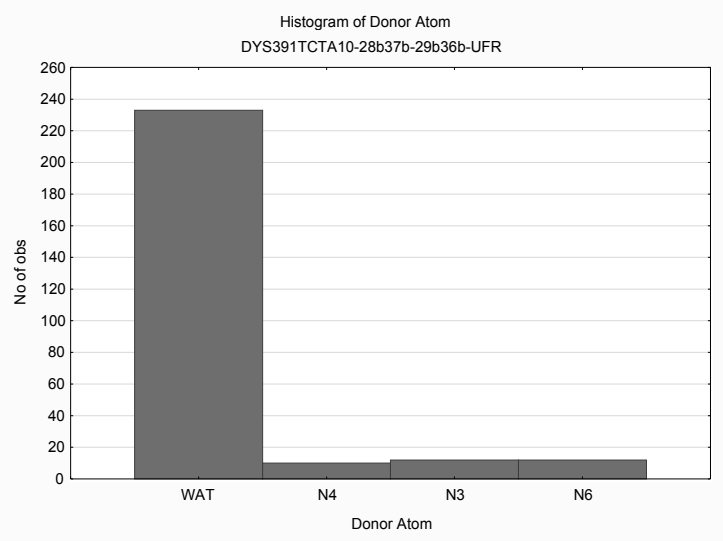

Histogram of Donor Residue
DYS391TCTA10-24b33b-25b32b-UFR
No of obs
Donor Residue


Histogram of Donor Atom
DYS391TCTA10-24b33b-25b32b-UFR
No of obs
0
20
40
60
80
100
120
140
160
180
200
220
240
260
WAT
N3
N6
N4
Donor Atom

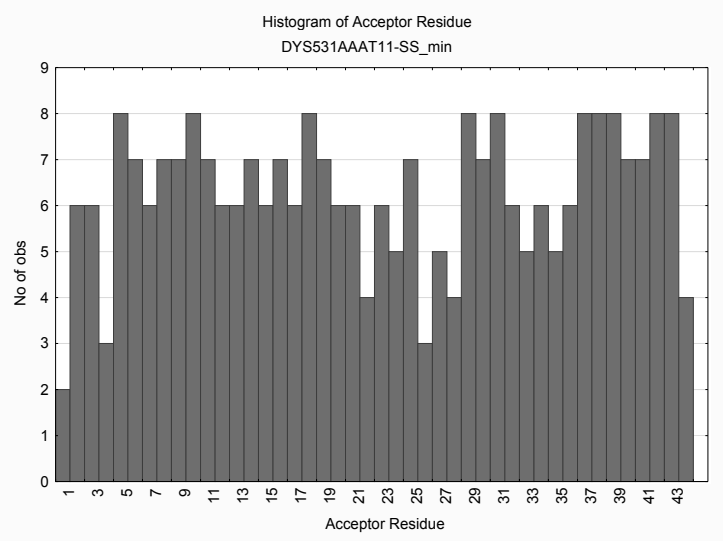

Histogram of Acceptor Residue
DYS391TCTA10-28b37b-29b36b-UFR
No of obs
0
1
2
3
4
5
6
7
8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
Acceptor Residue

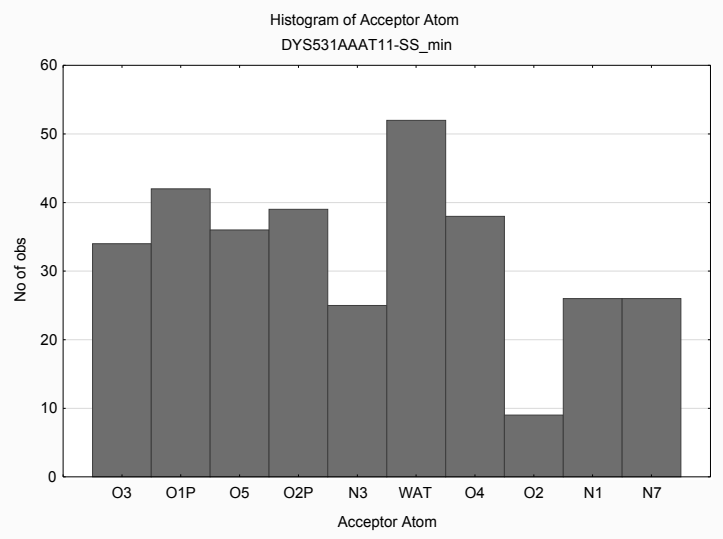

Histogram of Acceptor Atom
DYS391TCTA10-28b37b-29b36b-UFR
No of obs
0
5
10
15
20
25
30
35
40
45
O4
N3
O1P
WAT
O5
O2P
O2
N7
O3
N1
Acceptor Atom

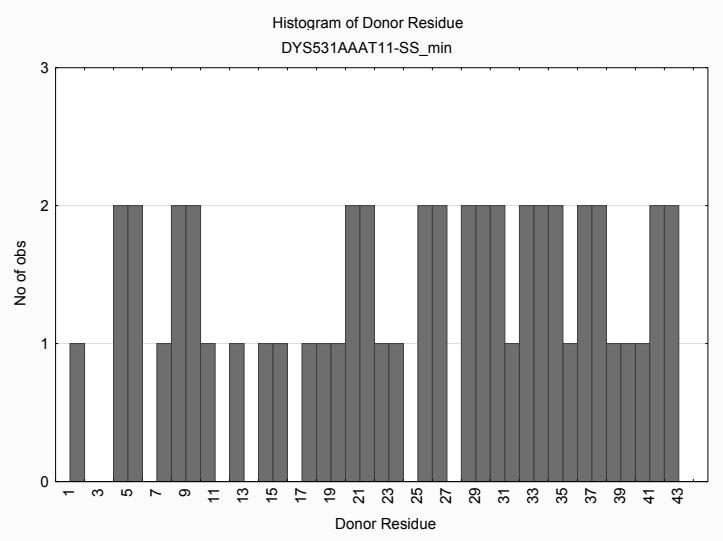

Histogram of Donor Residue
DYS391TCTA10-28b37b-29b36b-UFR
No of obs
Donor Residue

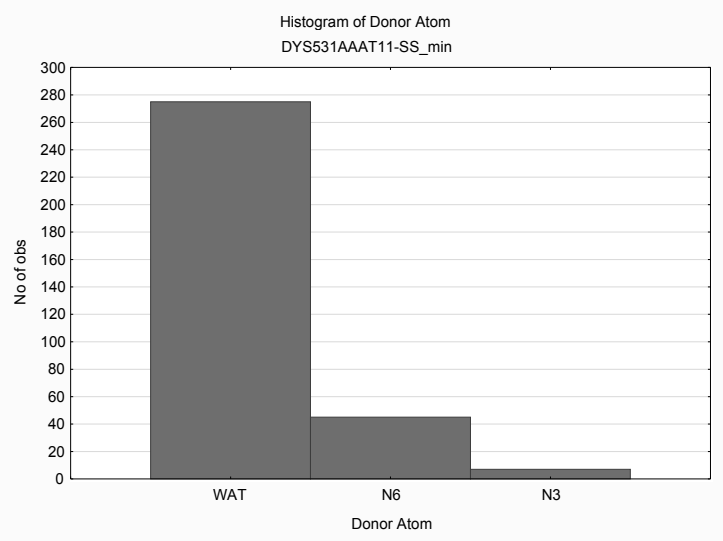

Histogram of Donor Atom
DYS391TCTA10-28b37b-29b36b-UFR
No of obs
WAT
N4
N3
N6
Donor Atom

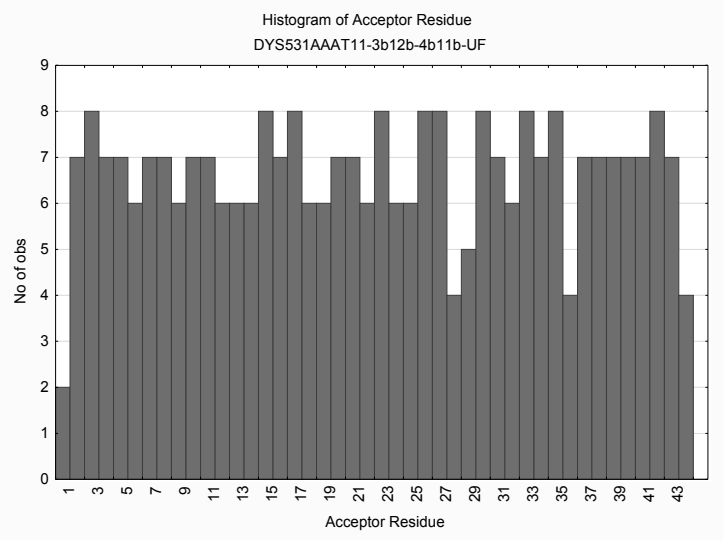
Graphs SI242-SI281: Histograms of water H-bonds (WAT; cut-off 3 Å) near nucleotides (acceptor and donor residue positions) and H-bonds between water and specific atom types (oxygen-phosphate, oxygen and nitrogen atoms), observed for the last 4 ns of molecular dynamics of DYS531 tested models.

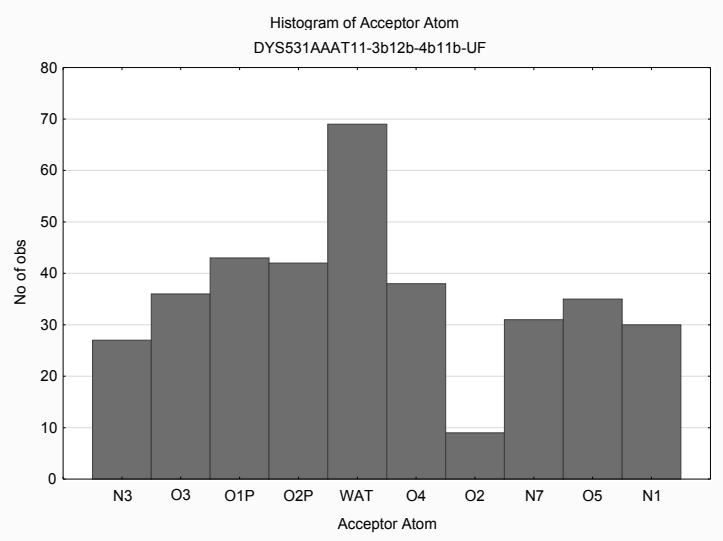


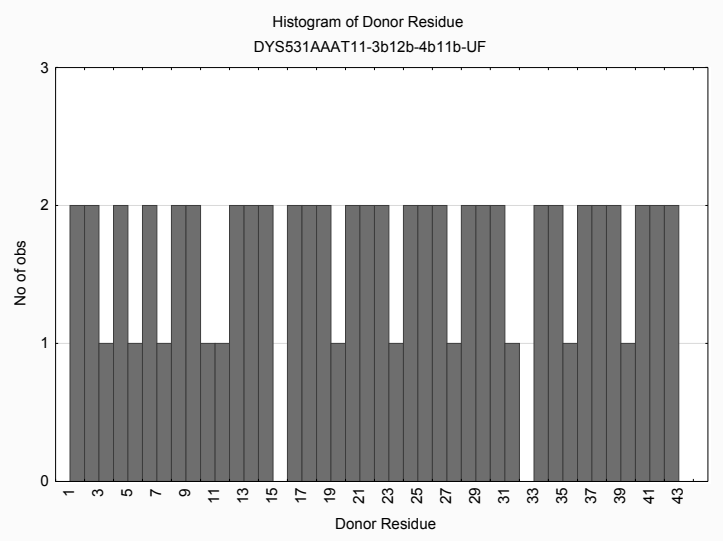


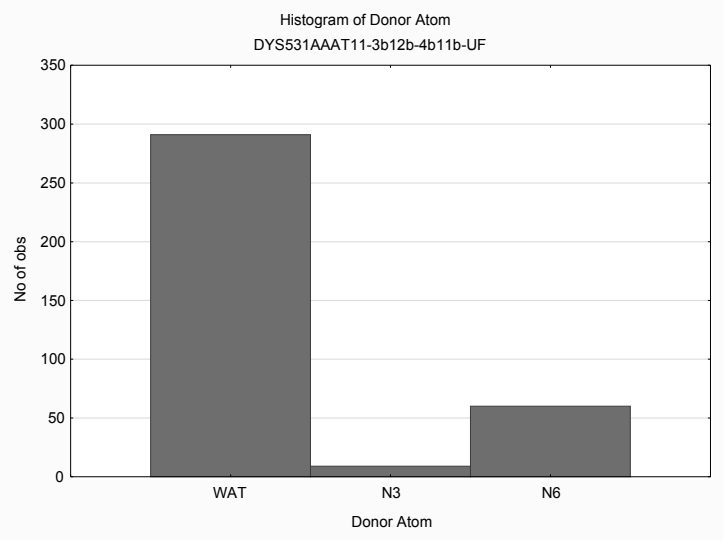

Histogram of Donor Residue
DYS531AAAT11-SS_min
No of obs
Donor Residue

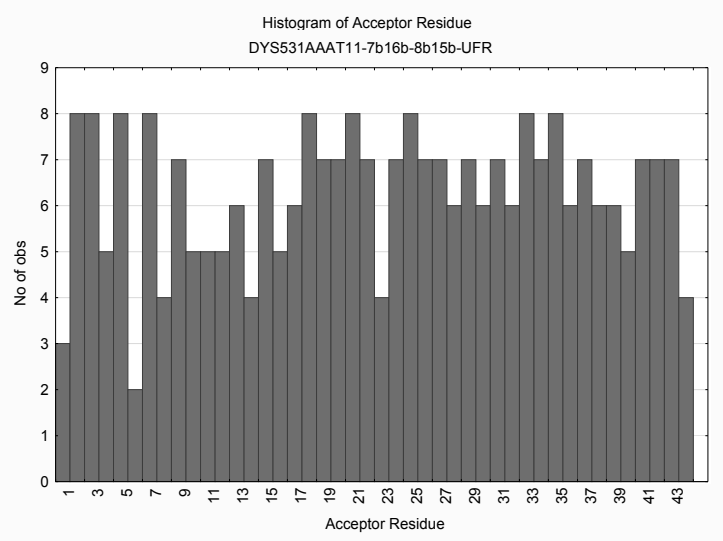

Histogram of Donor Atom
DYS531AAAT11-SS_min
No of obs
WAT
N6
N3
Donor Atom

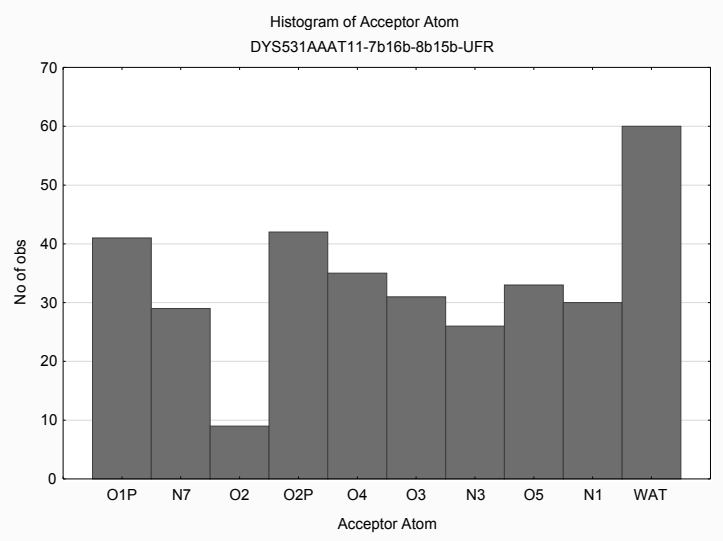

Histogram of Acceptor Residue
DYS531AAAT11-3b12b-4b11b-UF
No of obs
Acceptor Residue

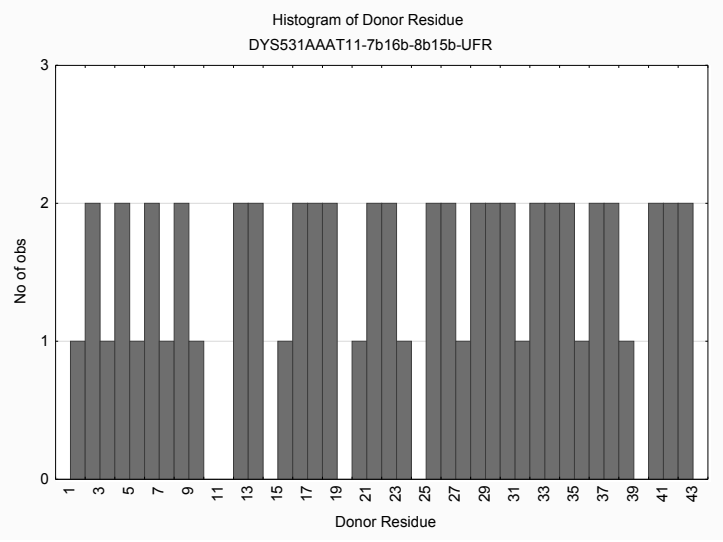

Histogram of Acceptor Atom
DYS531AAAT11-3b12b-4b11b-UF
No of obs
0
10
20
30
40
50
60
70
80
N3
O3
O1P
O2P
WAT
O4
O2
N7
O5
N1
Acceptor Atom

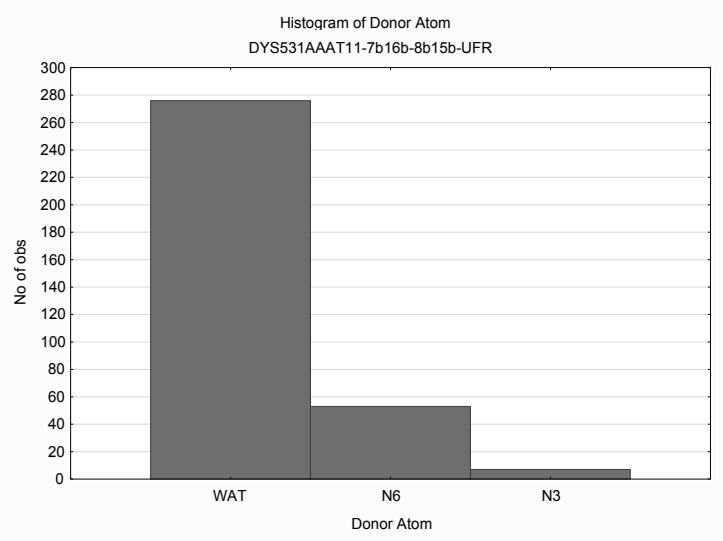

Histogram of Donor Residue
DYS531AAAT11-3b12b-4b11b-UF
No of obs
0
1
2
3
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
Donor Residue

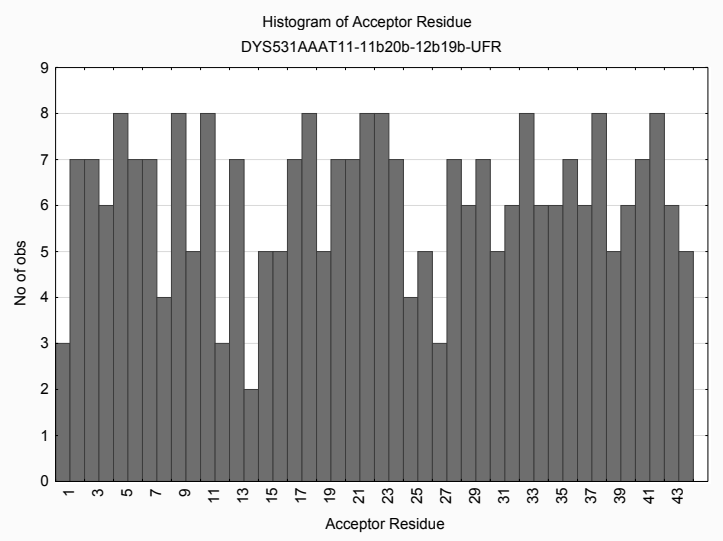

Histogram of Donor Atom
DYS531AAAT11-3b12b-4b11b-UF
No of obs
0
50
100
150
200
250
300
350
WAT
N3
N6
Donor Atom

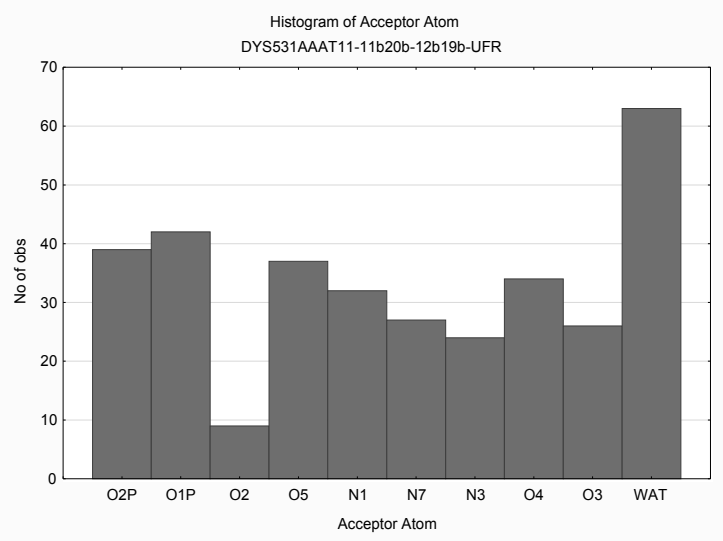

Histogram of Acceptor Residue
DYS531AAAT11-7b16b-8b15b-UFR
No of obs
Acceptor Residue

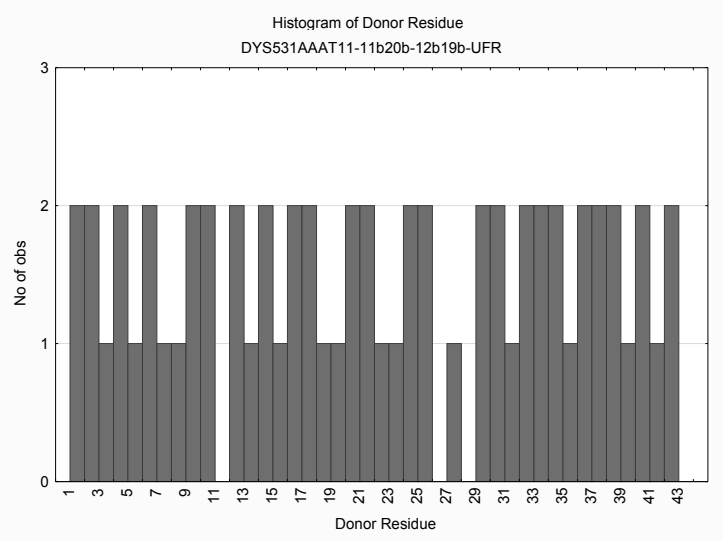

Histogram of Acceptor Atom
DYS531AAAT11-7b16b-8b15b-UFR
No of obs
O1P
N7
O2
O2P
O4
O3
N3
O5
N1
WAT
Acceptor Atom

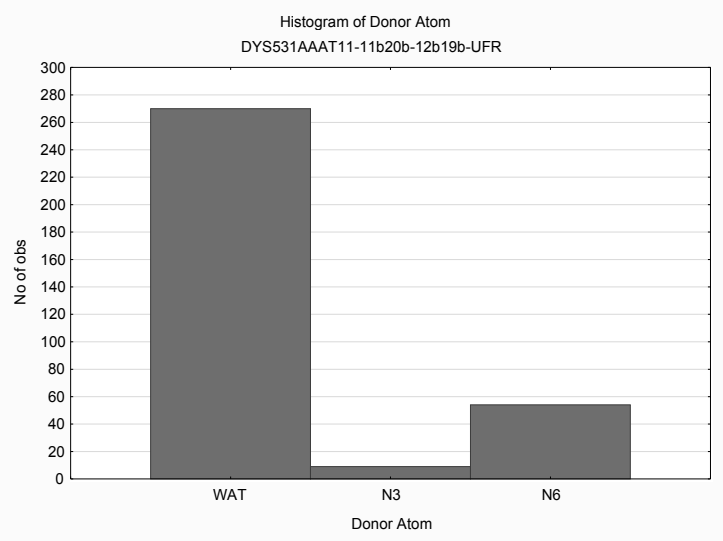

Histogram of Donor Residue
DYS531AAAT11-7b16b-8b15b-UFR
No of obs
Donor Residue

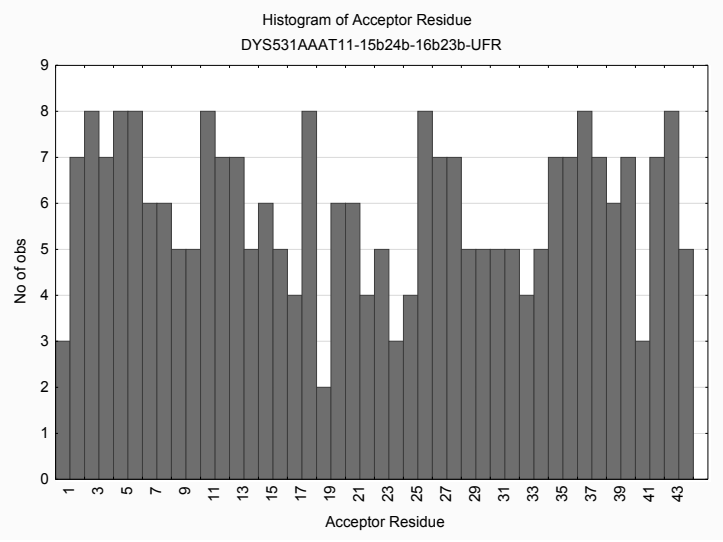

Histogram of Donor Atom
DYS531AAAT11-7b16b-8b15b-UFR
No of obs
Donor Atom
WAT
N6
N3

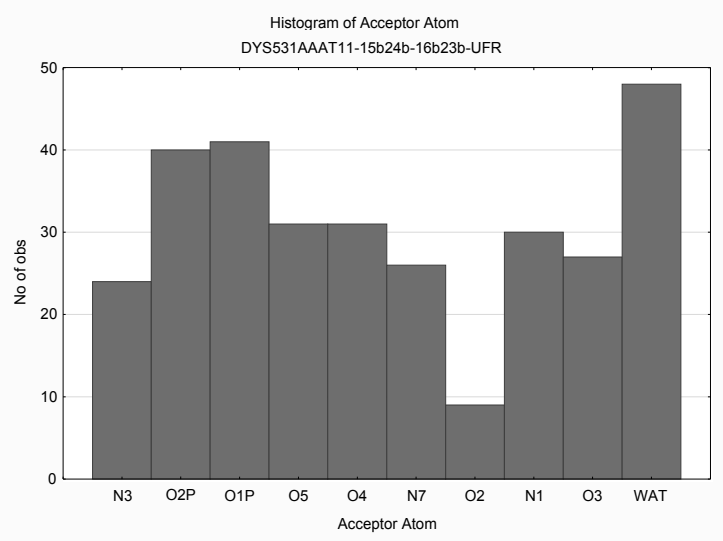

Histogram of Acceptor Residue
DYS531AAAT11-11b20b-12b19b-UFR
No of obs
Acceptor Residue

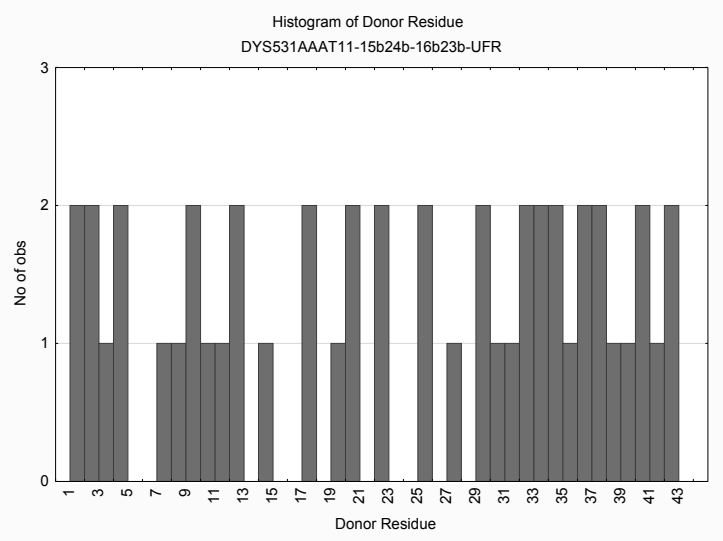

Histogram of Acceptor Atom
DYS531AAAT11-11b20b-12b19b-UFR
No of obs
Acceptor Atom
O2P
O1P
O2
O5
N1
N7
N3
O4
O3
WAT

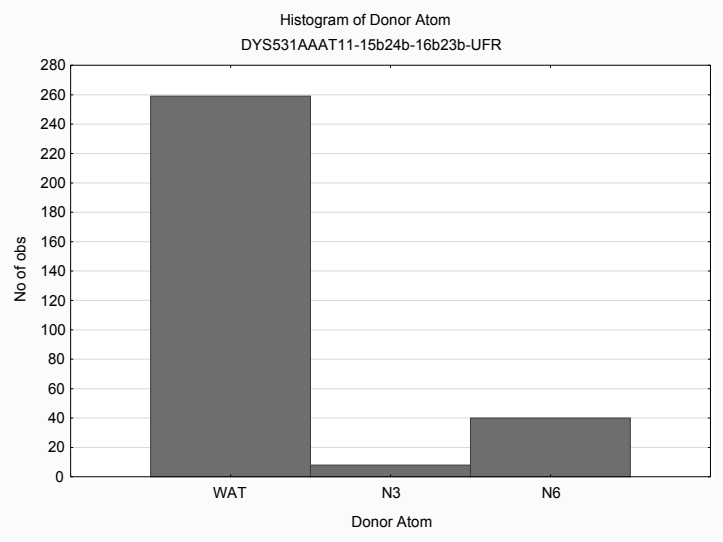

Histogram of Donor Residue
DYS531AAAT11-11b20b-12b19b-UFR
No of obs
Donor Residue

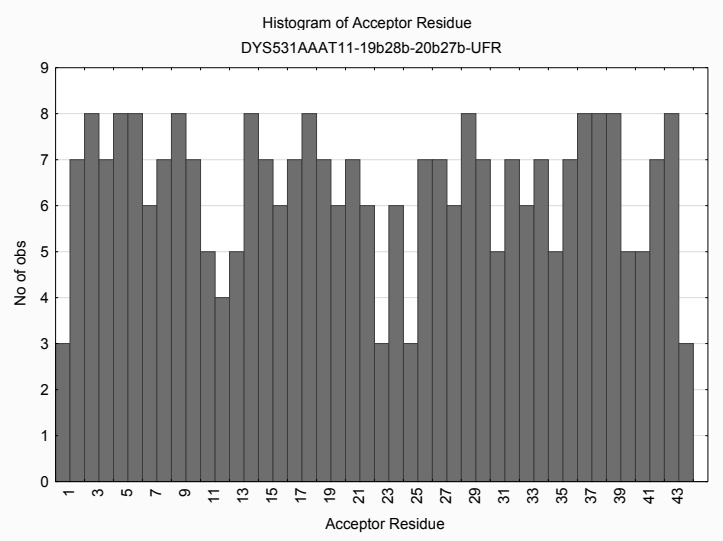

Histogram of Donor Atom
DYS531AAAT11-11b20b-12b19b-UFR
No of obs
WAT
N3
N6
Donor Atom

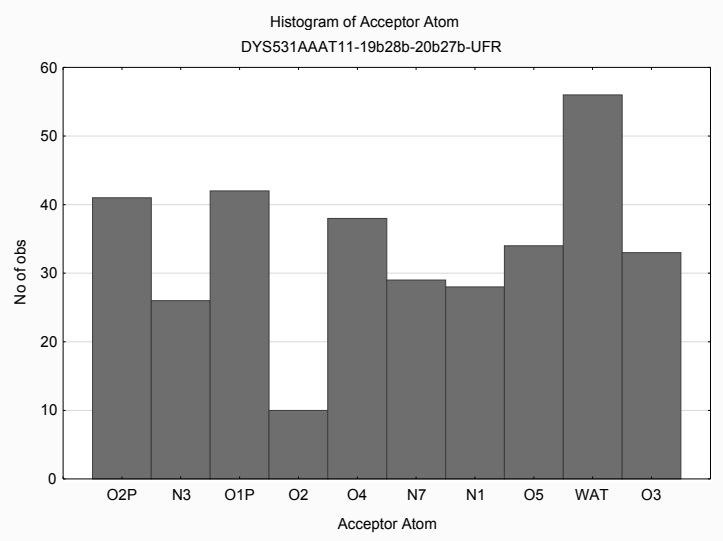

Histogram of Acceptor Residue
DYS531AAAT11-15b24b-16b23b-UFR
No of obs
Acceptor Residue

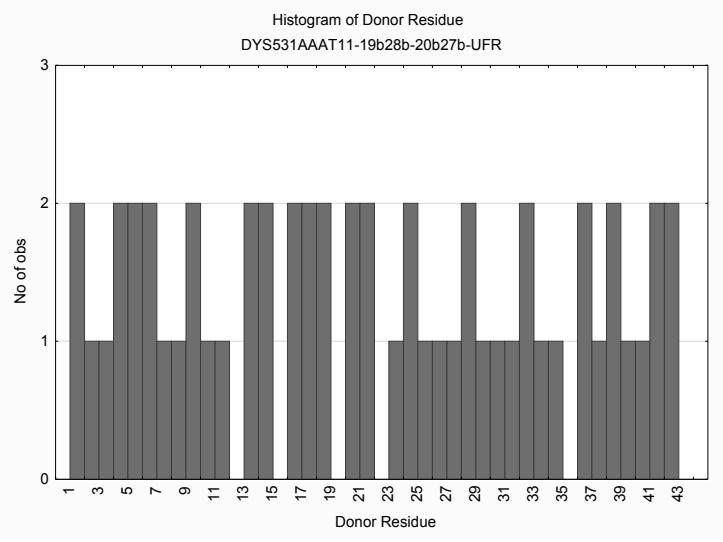

Histogram of Acceptor Atom
DYS531AAAT11-15b24b-16b23b-UFR
No of obs
0
10
20
30
40
50
N3
O2P
O1P
O5
O4
N7
O2
N1
O3
WAT
Acceptor Atom

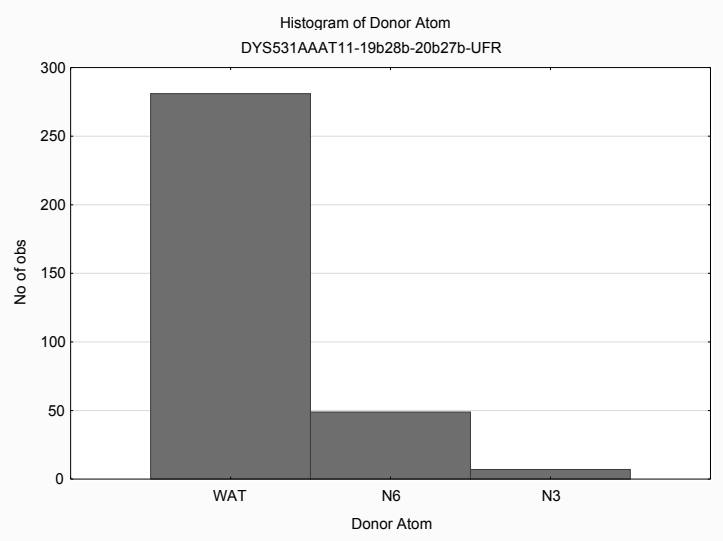

Histogram of Donor Residue
DYS531AAAT11-15b24b-16b23b-UFR
No of obs
0
1
2
3
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
Donor Residue

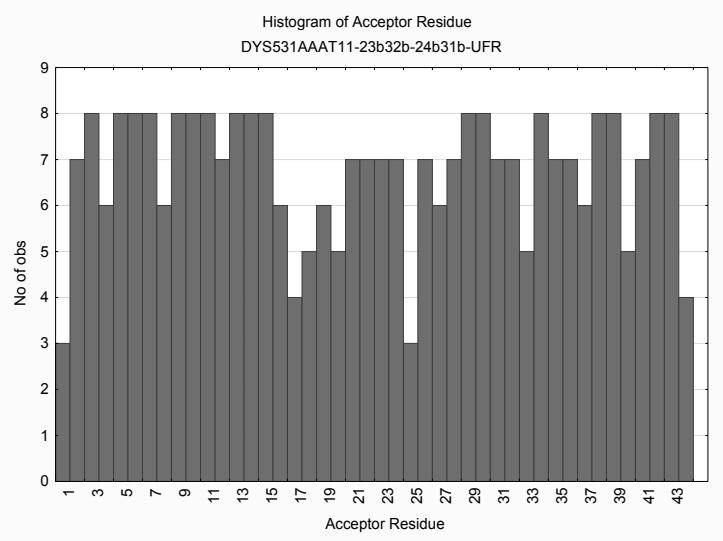

Histogram of Donor Atom
DYS531AAAT11-15b24b-16b23b-UFR
No of obs
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
WAT
N3
N6
Donor Atom

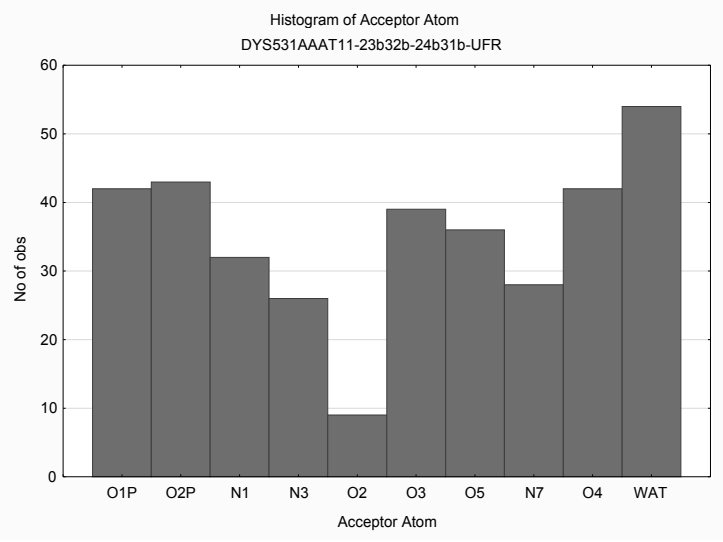

Histogram of Acceptor Residue
DYS531AAAT11-19b28b-20b27b-UFR
No of obs
Acceptor Residue

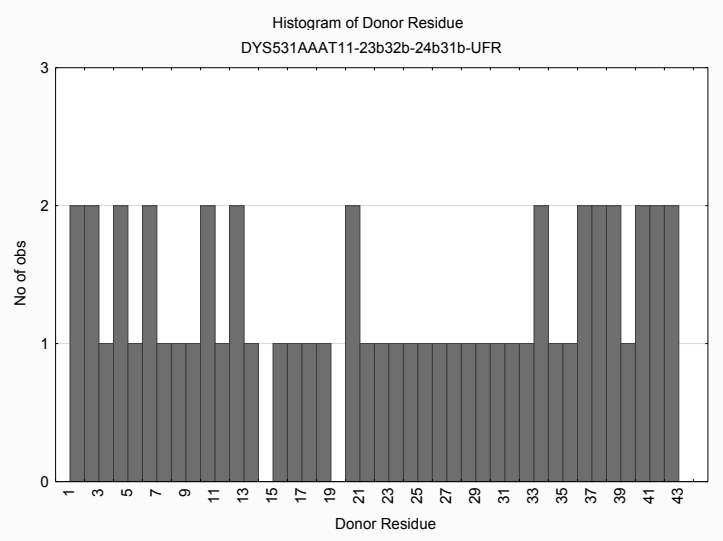

Histogram of Acceptor Atom
DYS531AAAT11-19b28b-20b27b-UFR
No of obs
O2P
N3
O1P
O2
O4
N7
N1
O5
WAT
O3
Acceptor Atom

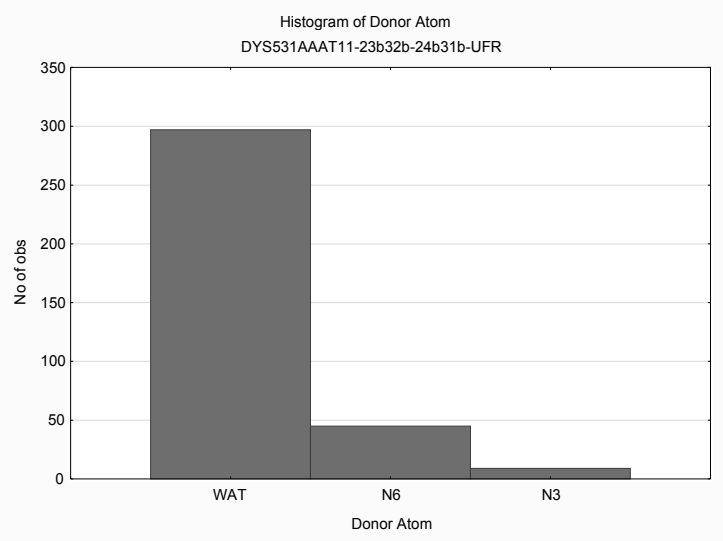

Histogram of Donor Residue
DYS531AAAT11-19b28b-20b27b-UFR
No of obs
Donor Residue

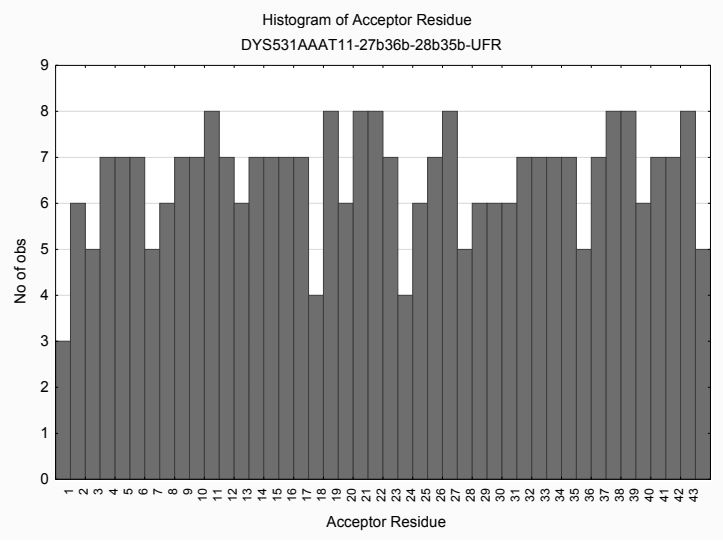

Histogram of Donor Atom
DYS531AAAT11-19b28b-20b27b-UFR
No of obs
300
250
200
150
100
50
0
WAT
N6
N3
Donor Atom

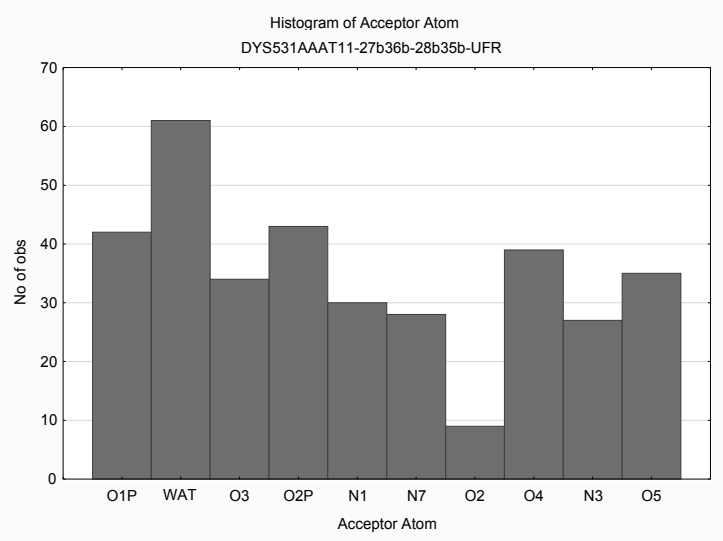

Histogram of Acceptor Residue
DYS531AAAT11-23b32b-24b31b-UFR
No of obs
9
8
7
6
5
4
3
2
1
0
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
41
43
Acceptor Residue

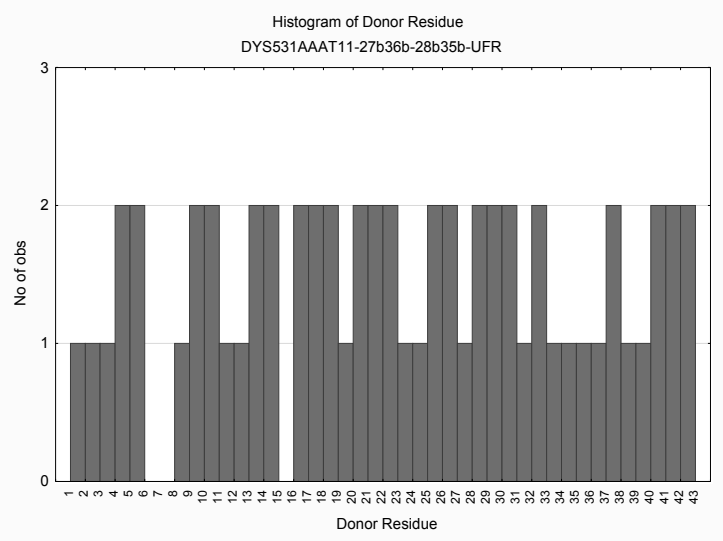

Histogram of Acceptor Atom
DYS531AAAT11-23b32b-24b31b-UFR
No of obs
60
50
40
30
20
10
0
O1P
O2P
N1
N3
O2
O3
O5
N7
O4
WAT
Acceptor Atom

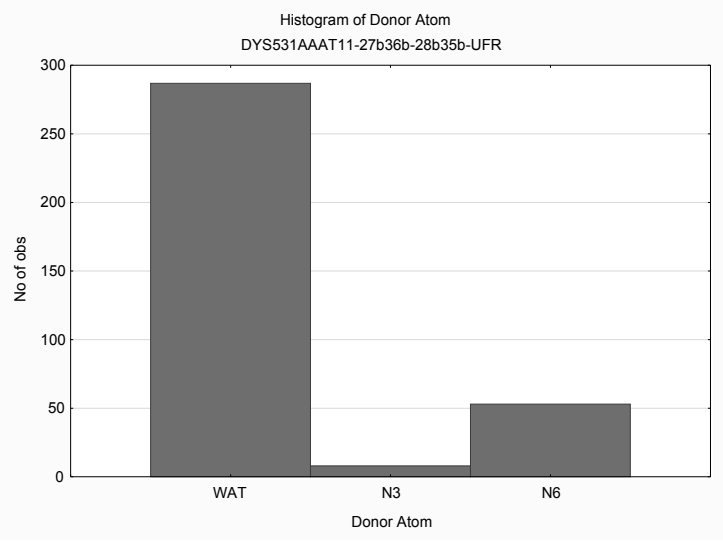

Histogram of Donor Residue
DYS531AAAT11-23b32b-24b31b-UFR
No of obs
Donor Residue

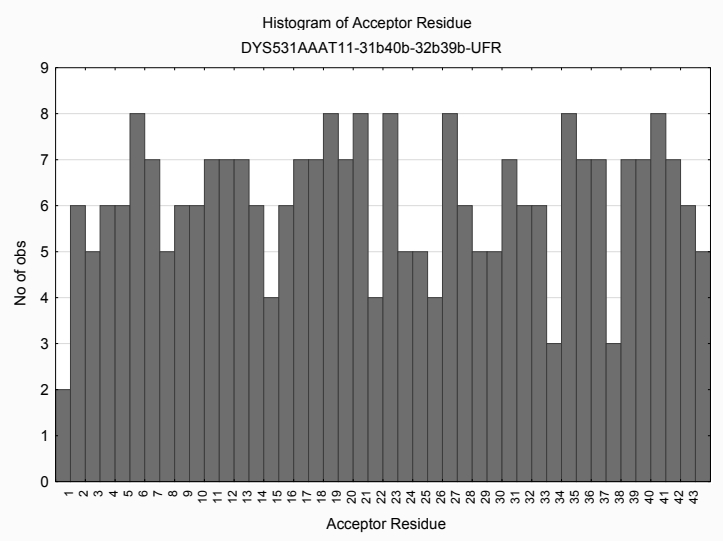

Histogram of Donor Atom
DYS531AAAT11-23b32b-24b31b-UFR
No of obs
WAT
N6
N3
Donor Atom

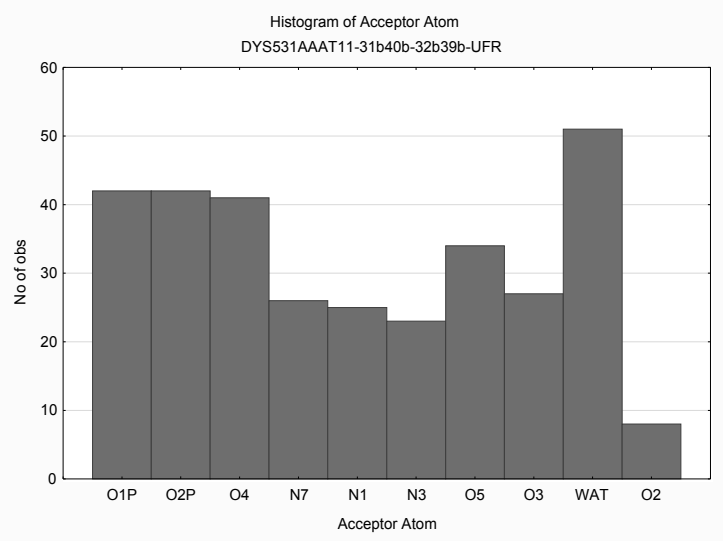

Histogram of Acceptor Residue
DYS531AAAT11-27b36b-28b35b-UFR
No of obs
Acceptor Residue

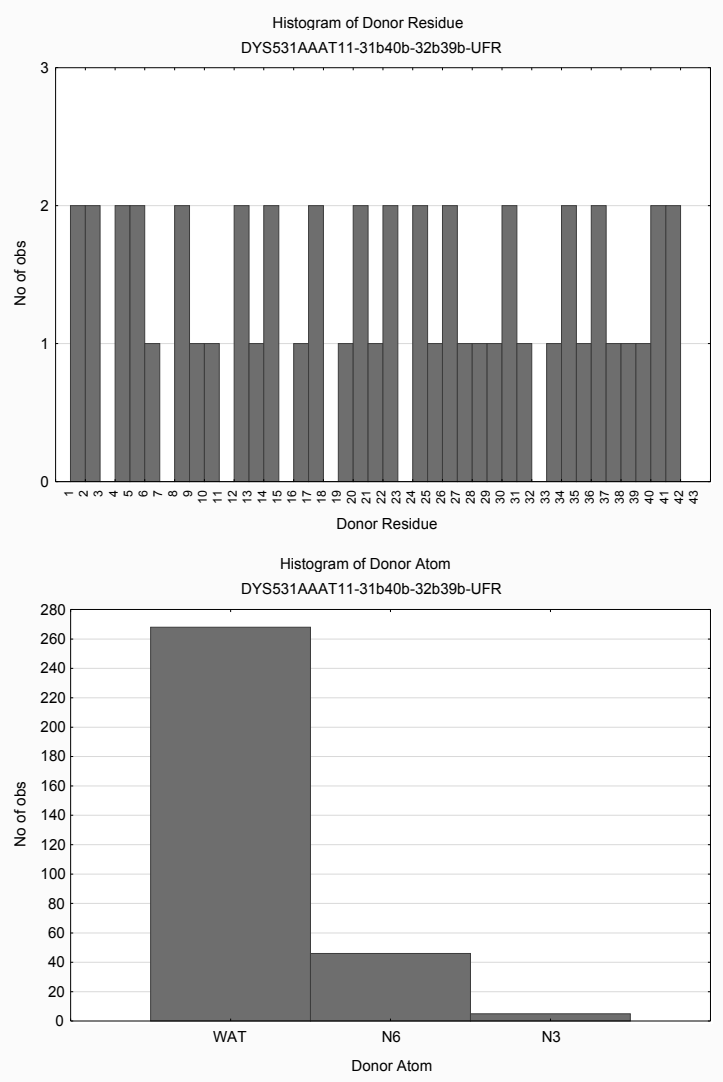

Histogram of Acceptor Atom
DYS531AAAT11-27b36b-28b35b-UFR
No of obs
0
10
20
30
40
50
60
70
O1P
WAT
O3
O2P
N1
N7
O2
O4
N3
O5
Acceptor Atom

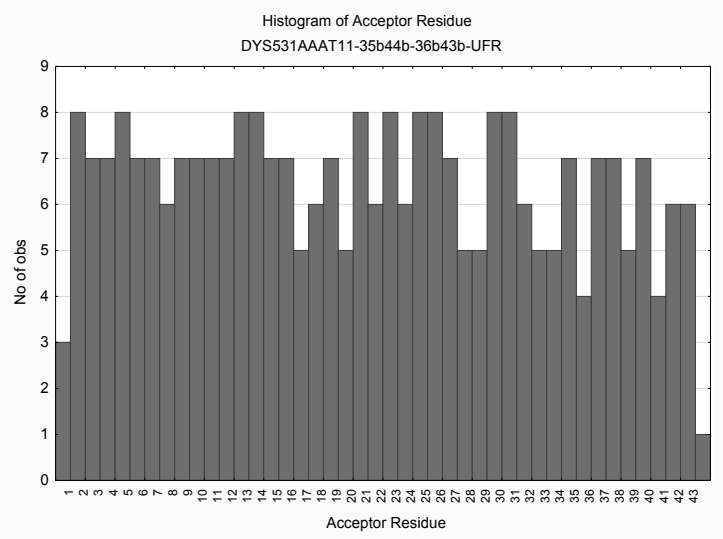

Histogram of Donor Residue
DYS531AAAT11-27b36b-28b35b-UFR
No of obs
0
1
2
3
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43
Donor Residue

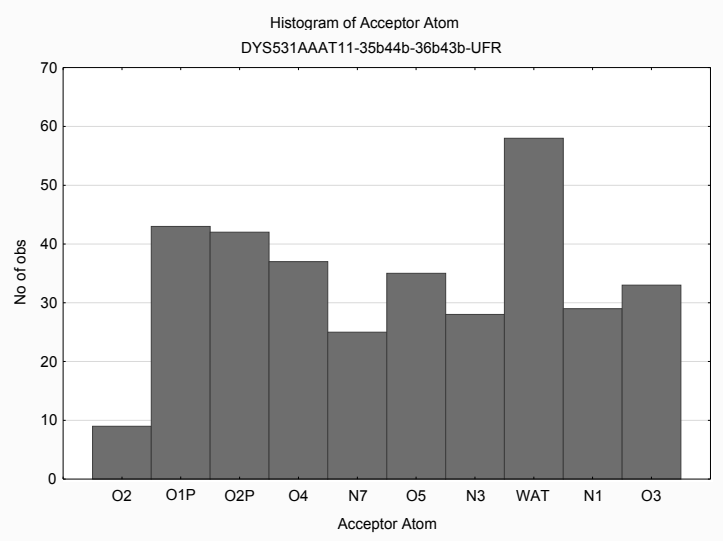

Histogram of Donor Atom
DYS531AAAT11-27b36b-28b35b-UFR
No of obs
0
50
100
150
200
250
300
WAT
N3
N6
Donor Atom

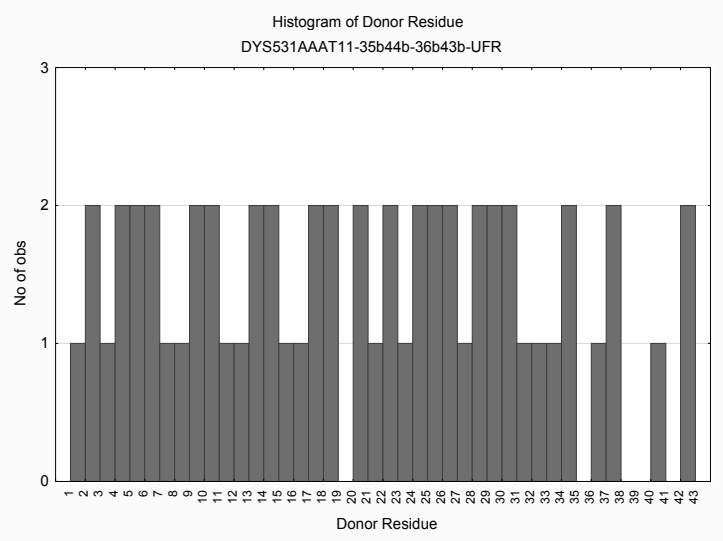

Histogram of Acceptor Residue
DYS531AAAT11-31b40b-32b39b-UFR
No of obs
Acceptor Residue


Histogram of Acceptor Atom
DYS531AAAT11-31b40b-32b39b-UFR
No of obs
O1P
O2P
O4
N7
N1
N3
O5
O3
WAT
O2
Acceptor Atom


Histogram of Donor Residue
DYS531AAAT11-31b40b-32b39b-UFR
No of obs
Donor Residue
Histogram of Donor Atom
DYS531AAAT11-31b40b-32b39b-UFR
No of obs
WAT
N6
N3
Donor Atom


Histogram of Acceptor Residue
DYS531AAAT11-35b44b-36b43b-UFR
No of obs
Acceptor Residue


Histogram of Acceptor Atom
DYS531AAAT11-35b44b-36b43b-UFR
No of obs
Acceptor Atom
O2
O1P
O2P
O4
N7
O5
N3
WAT
N1
O3


Histogram of Donor Residue
DYS531AAAT11-35b44b-36b43b-UFR
No of obs
Donor Residue

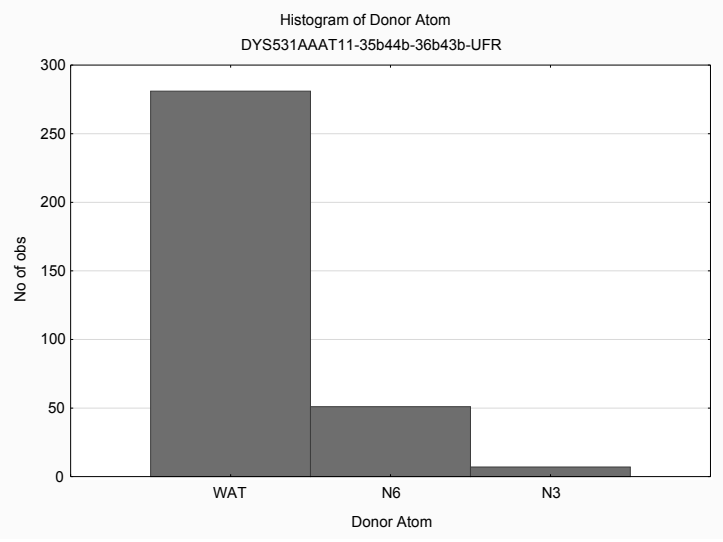

Histogram of Donor Atom
DYS531AAAT11-35b44b-36b43b-UFR
No of obs
Donor Atom
WAT
N6
N3

Table SI33: Number of local base-pairs detected for DYS19A, DYS391 and DYS531 molecular models (average PDB of last 4ns), as calculated by the X3DNA software[136].

Molecular Dynamics model	Number of base pairs detected	Mean
DYS19A-SS	5	5 Std:2.07
DYS19A-4b13b-5b12b-UF	10	
DYS19A-12b21b-13b20b-UFR	6	
DYS19A-20b29b-21b28b-UFR	5	
DYS19A-24b33b-25b32b-UFR	4	
DYS19A-28b37b-29b36b-UFR	5	
DYS19A-32b41b-33b40b-UFR	4	
DYS19A-36b45b-37b44b-UFR	2	
DYS19A-40b49b-41b48b-UFR	4	
DYS19A-44b53b-45b52b-UFR	3	
DYS19A-48b57b-49b56b-UFR	6	
DYS391TCTA10-SS	2	2.5 Std:0.74
DYS391TCTA10-4b13b-5b12b-UF	3	
DYS391TCTA10-8b17b-9b16b-UFR	2	
DYS391TCTA10-12b21b-13b20b-UFR	3	
DYS391TCTA10-16b25b-17b24b-UFR	2	
DYS391TCTA10-20b29b-21b28b-UFR	4	
DYS391TCTA10-24b33b-25b32b-UFR	2	
DYS391TCTA10-28b37b-29b36b-UFR	3	
DYS531AAAT11-SS	7	2 Std:1.87
DYS531AAAT11-3b12b-4b11b-UF	2	
DYS531AAAT11-7b16b-8b15b-UFR	2	
DYS531AAAT11-11b20b-12b19b-UFR	1	
DYS531AAAT11-15b24b-16b23b-UFR	1	
DYS531AAAT11-19b28b-20b27b-UFR	2	
DYS531AAAT11-23b32b-24b31b-UFR	2	
DYS531AAAT11-27b36b-28b35b-UFR	5	
DYS531AAAT11-31b40b-32b39b-UFR	3	
DYS531AAAT11-35b44b-36b43b-UFR	3	

11. Publication V: SPInDel - a multi-functional workbench for species identification using insertion/deletion variants

SPInDel: a multi-functional workbench for species identification using insertion/deletion variants

João Carneiro[1, 2], Filipe Pereira[1] and António Amorim[1, 2]

[1] Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), R. Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

[2] Faculty of Sciences, University of Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

Corresponding authors:

João Carneiro
(jcarneiro@ipatimup.pt)

Filipe Pereira
(fpereirapt@gmail.com)

IPATIMUP
R. Dr. Roberto Frias s/n, 4200-465 Porto, Portugal
Phone: +351 22 5570700
Fax: +351 22 5570799

Running title: Workbench for species identification using indels

Molecular Ecology Resources. 2012.

doi: 10.1111/1755-0998.12011

Abstract

The majority of the available methods for the molecular identification of species use pairwise sequence divergences between the query and reference sequences (DNA barcoding). The presence of multiple insertions and deletions (indels) in the target genomic regions is generally regarded as a problem, as it introduces ambiguities in sequence alignments. However, we have recently shown that a high level of species discrimination is attainable in all taxa of life simply by considering the length of hypervariable regions defined by indel variants. Each species is tagged with a numeric profile of fragment lengths – a true numeric barcode. In this study, we describe a multi-functional computational workbench (named SPInDel for SPecies Identification by Insertions/Deletions) to assist researchers using variable-length DNA sequences, and we demonstrate its applicability in molecular ecology. The SPInDel workbench provides a step-by-step environment for the alignment of target sequences, selection of informative hypervariable regions, design of PCR primers and the statistical validation of the species-identification process. In our test datasets, we were able to discriminate all species from two genera of frogs (Ansonia and Leptobrachium) inhabiting lowland rainforests and mountain regions of Southeast Asia and species from the most common genus of coral reef fishes (Apogon). Our method can complement conventional DNA barcoding systems when indels are common (e.g., in rRNA genes) without the required step of DNA sequencing. The executable files, source code, documentation and test datasets are freely available at http://www.portugene.com/SPInDel/SPInDel_webworkbench.html.

Keywords: species identification, insertions/deletions, numeric profiles, variable-length sequences, mtDNA, rRNA

Introduction

The identification of biological samples collected during ecological field work (e.g., wildlife species) is often a challenging task (Hebert et al. 2003; Steinke et al. 2005; Vences et al. 2005; Darling& Blum 2007; Pereira et al. 2008). In certain cases, scientists must investigate vestigial and highly degraded samples, such as carcasses, feces, bones, hair, teeth, eggshells, fur, feathers, stomach contents, seeds, and wood. Similarly, the emerging field of wildlife species identification is also dependent on the accurate identification of products made from protected animals (e.g., leather goods or medicinal powder) or crime scene evidence (e.g., bite wounds) (Coyle 2007; Rob Ogden et al. 2009; Amorim 2010; Linacre& Tobe 2011).

In certain circumstances, the only way to identify the species of origin of such vestigial samples is to apply molecular biology methods in the laboratory. However, many researchers conducting such investigations face limited laboratory equipment, and financial resources limit the use of DNA sequencing, microarrays or real-time PCR (McManus& Bowles 1996; Darling& Blum 2007; Pereira et al. 2008; Wells& Stevens 2008; Alacs et al. 2010). We have recently shown that a high level of species discrimination is attainable in all taxa of life simply by determining and combining the length of hypervariable regions with indel variants (Pereira et al. 2010). The numeric profiles that identify each species can be assessed using diverse genotyping platforms, including those requiring low-cost equipment and reagents (e.g., conventional agarose or polyacrylamide gels). Our method enables inter-laboratory comparison, providing a means to standardize methodologies. In our datasets, the levels of intraspecies variation are comparable to those detected by sequencing analyses (Pereira et al. 2010). Our method also permits the identification of species from admixtures and is appropriate for low-quantity and/or degraded DNA samples (very short amplicons can be used, for instance, shorter than 100 bp).

It has been shown that genomic regions with multiple indels can be used for species-identification procedures in animals [rRNA gene sequences (Steinke et al. 2005; Vences et al. 2005)], plants [chloroplast trnL (UAA) intron (Taberlet et al. 2007)], fungi [ITS; (Zinger et al. 2008)] and bacteria [rRNA; (Sogin et al. 2006)], with the same efficiency as using mitochondrial cytochrome oxidase subunit I (cox1). Moreover, because indels are less prone to recurrent and back mutations, the probability of misclassifications is greatly reduced. Several software tools are now available for indel detection in deep-sequencing data (Young& Healy 2003; Neuman et al. 2012), providing the necessary genetic information for the development of new indel-based identification systems. The SPInDel computational workbench described here can be used with sequence data from any genomic region and is a useful tool to help researchers in all steps of the species identification workflow.

Features and basic usage

The SPInDel computational platform (Figure 1) was designed to facilitate the planning and management of projects for the analysis of indel variability in sequence datasets. It was built using the high-level object-oriented programming language Python (Python 2.6, freely available at www.python.org) and other third-party packages (Supplementary Information S1 and S2). A step-by-step description of the procedures for using the SPInDel workbench and the theoretical background on SPInDel calculations are presented in Supplementary Information S3 and Figure 2.

Multiple sequence alignments

A FASTA-formatted file or an SQLite database with aligned DNA or RNA sequences (haplotypic data) can be uploaded to the SPInDel SQLite database (Supplementary Information S4). Sequence re-alignments can also be performed in the workbench with the PyCogent progressive alignment algorithm. The user can also select among different nucleotide substitution models (JC69, F81, HKY85 and GTR) and different rates of occurrence of indels. The main window plots an identity value for each nucleotide position by estimating the frequency of the most common nucleotide in the aligned sequences. This feature allows easy identification of conserved regions (highest conservation and lowest conservation are represented in green and red, respectively) that can be chosen directly in the alignment window using column selection.

Numerical profiles of fragment lengths

Conserved regions of multiple sequence alignments are used to delimit the target segments with indels ("SPInDel hypervariable regions"). The combination of sequence lengths on different SPInDel hypervariable regions produces unique numeric combinations for each sequence or group of identical sequences (a "SPInDel profile"). A function that computes the discriminatory power of all combinations of hypervariable regions can be used to identify the minimum number of regions for an accurate identification. The algorithm generates n-combinations without repetition, which are subsets of n distinct elements of the set of all possible regions. For each n-combination, the numbers of shared profiles (Nsp) and different profiles (Ndp) are displayed in tables and graphs. To avoid the design of complementary PCR primers at the same location, the 'multiplex PCR option' retrieves only n-combinations that do not share conserved regions. Diverse distance measures are implemented by the use of in-house developed Python algorithms.

The identity of a numeric profile of unknown origin can be predicted with the 'Search profile' function by a k-nearest-neighbor method using a database of known profiles built with the SQLite3 Python SQL interface. This discrete metric was implemented using BioPython and an in-house developed function for the discrete distance metric. To test the accuracy of the classification, we implemented a leave-one-out cross validation using profiles from known species.

Step-by-Step Tutorial

Here, we describe how to perform a basic SPInDel workbench analysis using the genera Ansonia as an example. The following steps can be adopted by the user (with small modifications) for other taxonomic groups:

1. We retrieved mitochondrial rRNA gene sequences for Ansonia from the NCBI Entrez Nucleotide database.
2. We then randomly selected one representative of each species in Ansonia and performed a multiple sequence alignment for the 22 mitochondrial rRNA gene sequences obtained using freely available software.
3. A new SPInDel project was created using the 'New project' function in the top menu 'File,' and the Ansonia FASTA-formatted alignment was imported to the SPInDel workbench.
4. We identified 7 conserved regions in the multiple sequence alignment (top window), using as a guide the identity values (green regions with values higher than 0.95) in the bottom window.
5. The numeric profiles for each species were calculated with the 'Calculate profiles' function. The profiles were analysed using diverse statistical methods (e.g., Region by Region, Mismatch distribution and Combinations functions). The frequency of species-specific profiles is 1.00, indicating that all species have a unique SPInDel profile.
6. The UPGMA tree and the principal component analysis (PCA) were used to display the overall relationship among the numeric profiles.
7. We estimated the minimum number of regions for a complete discrimination of Ansonia species using the 'Combinations' function (in this case, 3 hypervariable regions were sufficient).
8. We tested whether the PCR primers' properties were in accordance with the Oligocalc (Supplementary Information S3) values for optimized PCR (standard or multiplex).
9. The numerical profiles of each hypervariable region (hypervariable region length) and PCR primers' properties were exported using the SPInDel exporter tools (to Excel using comma separated values files). This information could be used for the development of a laboratorial procedure for identification of Ansonia species.

The application of the SPInDel concept to taxonomic groups of ecological value

We have analysed 3 genera (Ansonia, Leptobrachium and Apogon) as representatives of two taxonomic groups (Amphibia and Actinopterygii) of great importance for ecological genetics studies. These genera were selected because they offered the

greatest number of available mitochondrial rRNA gene sequences in each group. The rRNA genes of mtDNA are particularly useful as targets of our approach due to the presence of multiple indels and highly conserved domains. A total of 173 sequences from the mtDNA region of the 12s rRNA, tRNA-Val and 16s rRNA genes (74, 61 and 38 sequences from Ansonia, Leptobrachium and Apogon, respectively) were initially retrieved from the NCBI Entrez Nucleotide database (http://www.ncbi.nlm.nih.gov). We then randomly selected one representative of each species in each genus and analysed the datasets in the SPinDel workbench (Table 1).

All species from the 3 genera have different profiles, and the average numbers of pairwise differences among hypervariable regions are 4.37 (Ansonia), 3.04 (Leptobrachium) and 5.44 (Apogon) (Table 1). The discrimination of all species is possible with only 3 (Ansonia) and 4 (Leptobranchium and Apogon) hypervariable regions. In general, our method was able to unambiguously discriminate closely related species in well-supported monophyletic clades (Mabuchi et al. 2006; Matsui et al. 2010a; Matsui et al. 2010b). To test the level of intraspecies variability, we ran a dataset including all sequences initially retrieved (i.e., including different sequences from the same species) (Supplementary Information S5). In Ansonia, only one shared profile was found between individuals of different species (*A. platysoma and A. minuta*). The remaining 16 profiles were found in individuals of the same species. Strikingly, two different profiles were found among *A. spinulifer*, in agreement with the previously observed maximum likelihood and Bayesian phylogenies. A similar result was found in the complete Leptobranchium dataset, with only one case of two species sharing the same profiles. The analysis of Apogon, the most species-rich genus of the reef fish family Apogonidae, revealed that all available species have unique profiles (Supplementary Information S5). Our method was able to clearly discriminate species with similar phylogenetic and morphological features (Mabuchi et al. 2006). Nevertheless, the identification of species in other taxonomic groups should be preceded by a detailed analysis of intra- and interspecies diversity levels, as recommended for any identification system.

The SPInDel workbench also includes an extensive database with more than 1,800 species-specific profiles from 18 major taxonomic groups. These groups include several critically endangered species, whose profiles can be used to design specific laboratory methods for their detection (e.g., the Bactrian camel, Sumatran orang-utan, kakapo, blue whale, Asian elephant, giant panda, tiger and bonobo). These data might be useful for improving the high-throughput analysis of samples in wildlife investigations, and the SPInDel workbench described here has all of the required tools to facilitate such procedures.

Acknowledgements

This study was supported by a research grant SFRH/BPD/44637/2008 to FP and the research project PTDC/CVT/100881/2008 from the Portuguese Foundation for Science and Technology (FCT). IPATIMUP is an Associate Laboratory of the Portuguese Ministry of Science, Technology and Higher Education and is partially supported by FCT.

Table 1: General description of standard SPInDel profiles in 3 test datasets: *Ansonia*, *Apogon and Leptobrachium*.

Taxonomic group	Number of sequences (*N*)	Number of conserved regions	Number of hypervariable regions (*n*)	Average number of pairwise differences ($\bar{p}_n^G$)	Average number of pairwise differences per hypervariable region	Number of species-specific profiles (N_{sp})	Frequency of species-specific profiles (f_n^G)	Number of species-shared profiles	Number of minimum hypervariable regions for discrimination of all species
Eukaryotes									
Ansonia (Amphibia)	22	7	6	4.37	0.73	22	1	0	3
Apogon (Actinopterygii)	36	9	8	5.44	0.68	36	1	0	4
Leptobrachium (Amphibia)	17	6	5	3.04	0.61	17	1	0	4

Figures

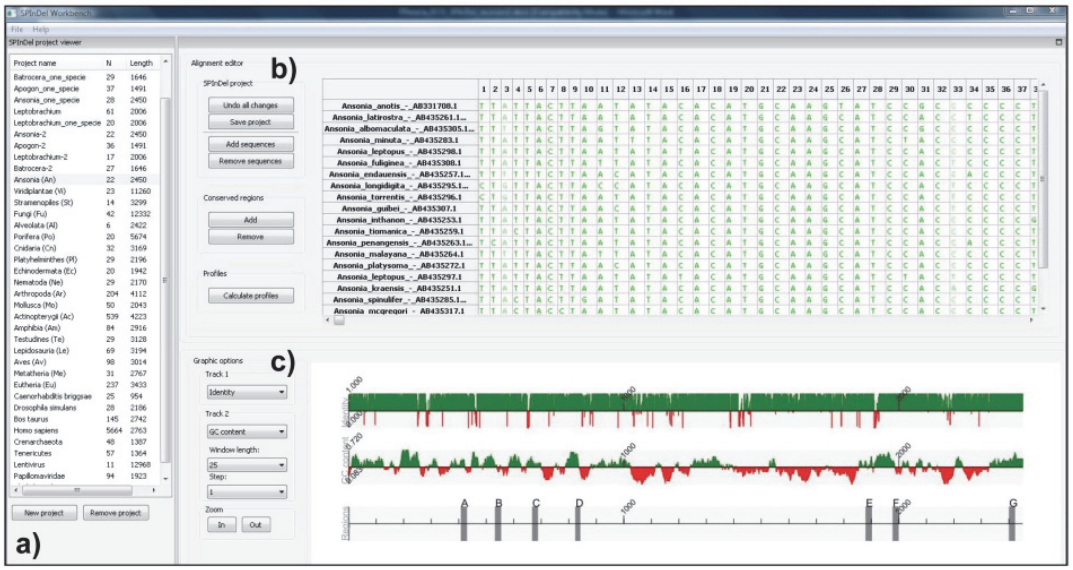


Figure 1: Main frame of the SPInDel workbench: a) selection box of SPInDel projects, b) sequence alignment viewer and c) alignment identity and GC content tracks.

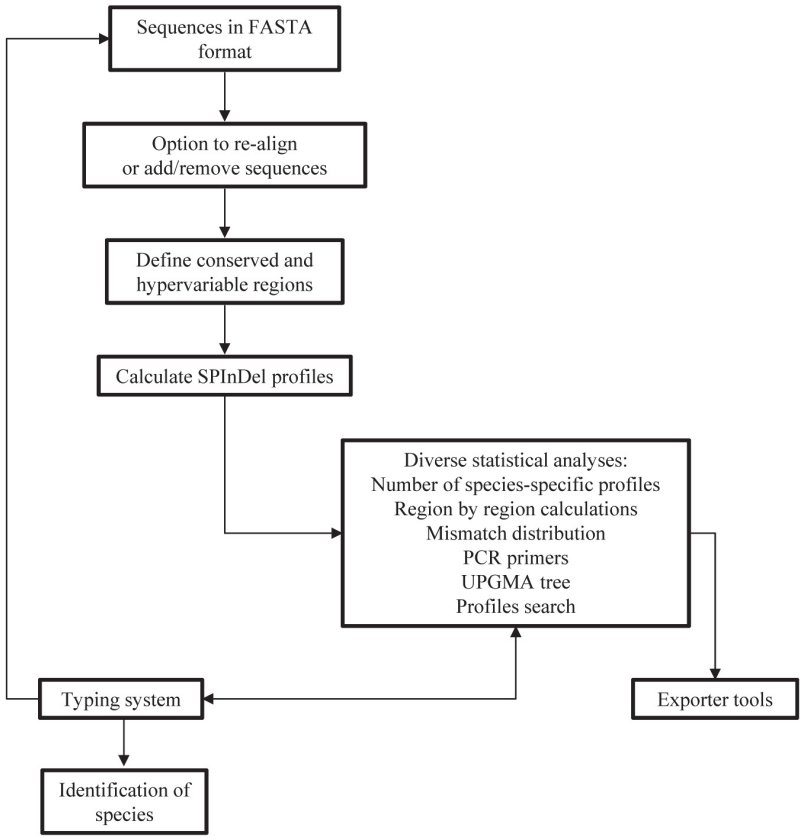


Figure 2: Flowchart of SPInDel workbench information processing for species identification.

Data accessibility
DNA sequences for analysis: Supplementary Information file.
The source code, documentation, test datasets with results and executable files for Windows and Linux are freely available at http://www.portugene.com/SPInDel/SPInDel_webworkbench.html.

Supporting Information
Additional supporting information may be found in the online version of this article.
Supplementary information: SPInDel workbench version 1.1 reference manual.

References

1. McManus DP, Bowles J (1996) Molecular genetic approaches to parasite identification: their value in diagnostic parasitology and systematics. Int J Parasitol 26, 687-704.
2. Hebert PD, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proc Biol Sci 270, 313-321.
3. Young ND, Healy J (2003) GapCoder automates the use of indel characters in phylogenetic analysis. BMC Bioinformatics 4, 6.
4. Steinke D, Vences M, Salzburger W, Meyer A (2005) TaxI: a software tool for DNA barcoding using distance methods. Philos Trans R Soc Lond B Biol Sci 360, 1975-1980.
5. Vences M, Thomas M, van der Meijden A, Chiari Y, Vieites DR (2005) Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. Front Zool 2, 5.
6. Mabuchi K, Okuda N, Nishida M (2006) Molecular phylogeny and stripe pattern evolution in the cardinalfish genus Apogon. Mol Phylogenet Evol 38, 90-99.
7. Sogin ML, Morrison HG, Huber JA, et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A 103, 12115-12120.
8. Coyle HM (2007) Nonhuman DNA Typing Theory and Casework Applications (ed. Coyle HM).
9. Darling JA, Blum MJ (2007) DNA-based methods for monitoring invasive species: a review and prospectus Biological Invasions 9, 751–765.
10. Taberlet P, Coissac E, Pompanon F, et al. (2007) Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. Nucleic Acids Res 35, e14.
11. Pereira F, Carneiro J, Amorim A (2008) Identification of species with DNA-based technology: current progress and challenges. Recent Pat DNA Gene Seq 2, 187-199.
12. Wells JD, Stevens JR (2008) Application of DNA-based methods in forensic entomology. Annu Rev Entomol 53, 103-120.
13. Zinger L, Gury J, Alibeu O, et al. (2008) CE-SSCP and CE-FLA, simple and high-throughput alternatives for fungal diversity studies. J Microbiol Methods 72, 42-53.
14. Rob Ogden, Nick Dawnay, McEwing R (2009) Wildlife DNA forensics—bridging the gap between conservation genetics and law enforcement. Endangered Species Research 9, 179-195.
15. Alacs EA, Georges A, FitzSimmons NN, Robertson J (2010) DNA detective: a review of molecular approaches to wildlife forensics. Forensic Sci Med Pathol 6, 180-194.
16. Amorim A (2010) Introduction to the Special Issue on Forensic Genetics: Non-Human DNA. The Open Forensic Science Journal Volume 3.
17. Matsui M, Hamidy A, Murphy RW, et al. (2010a) Phylogenetic relationships of megophryid frogs of the genus Leptobrachium (Amphibia, Anura) as revealed by mtDNA gene sequences. Mol Phylogenet Evol 56, 259-272.
18. Matsui M, Tominaga A, Liu W, et al. (2010b) Phylogenetic relationships of Ansonia from Southeast Asia inferred from mitochondrial DNA sequences: systematic and biogeographic implications (Anura: Bufonidae). Mol Phylogenet Evol 54, 561-570.
19. Pereira F, Carneiro J, Matthiesen R, et al. (2010) Identification of species by multiplex analysis of variable-length sequences. Nucleic Acids Res 38, e203.
20. Linacre A, Tobe SS (2011) An overview to the investigative approach to species testing in wildlife forensic science. Investig Genet 2, 2.
21. Neuman JA, Isakov O, Shomron N (2012) Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. Briefings in Bioinformatics 13.

12. General Discussion

12.1. Gene Families: The Metallothioneins (model a)

The main topics of this thesis are the analyses of molecular structural information of specific non-coding regions and the correlation of these data with the coding genome. First, a gene family that evolved from a series of duplications during evolution was studied. MTs (model a) are important metal-binding proteins involved in homeostasis and the transport of essential metals [35, 36]. In *Mus musculus* all MT genes are functional and have one copy. The human *MT2, MT3* and *MT4* are represented by one copy but there is a tandemly duplicated array of the human MT1, comprising eight active genes (*MT1A* to *MT1J, MT1L, MT1M* and *MT1X*) and five pseudogenes (*MT1L, MT1J, MT1D, MT1C and MT1I*). *Homo sapiens MT1* and *MT2* are ubiquitous expressed in all tissues, which is in agreement with their function as housekeeping genes that regulate heavy metal homeostasis in every cell. Human expression of *MT3* has been almost exclusively related to brain tissues, but our results demonstrated that it has a ubiquitous expression. Concerning *MT4*, although the degree of conservation of *MT4* was high between humans and mice, the presence of prohibitive residues in human sequence indicated that this protein might no longer be functional. In that sense, considering that inactivation of *MT4* occurred in some individuals, the role of the protein in epithelium where it is expressed was hypothesized to be assumed by *MT1* and/or *MT2*. *MT1* duplicates are good candidates to assume the function of the inactive *MT4* gene since they have similar metal binding properties, cellular specificity expression, and resistance to Cd and Cu. The analysis of phylogenetic data and expression profiles was determinant to establish the status of active/non-active genes in the MT family. Specific mutational change that occurs in duplicated copies can determine the non-coding status of the DNA regions that are translated into proteins. This analysis can be applied to other important gene families to access the functional relevance of the expressed proteins and the real impact of loss of function in genes (pseudogenisation). The observed conversion that can occur between coding and non-coding region by mutational steps was the support to infer that non-coding regions can share some features with coding regions. In this case the background DNA sequences were shared for both coding (genes) and non-coding sequences (pseudogenes), even though they change through evolutionary time. This was demonstrated in the phylogenetic analysis.

A deep understanding of the relationships between coding and non-coding regions was studied in models c and d. These non-transcribed non-coding regions (control region

mtDNA and STRs) can adopt non-B conformations that were analysed in order to see their influence in protein-coding genome.

12.2. NAD Pathway Relevant Genes: NAMPT and PNC (model b).

The pseudogenisation process can convert a protein coding region in a non-transcribed non-coding region (e.g., 5' UTR mutations that impair transcription initiation, mutations that affect the termination codon and cause diseases, mutations at 5' UTR secondary structures) [139, 140], but the consequences of mutations that do not influence transcriptional processes must be studied by protein models (model b). The gene that codes for a specific protein can be expressed but the protein can lose the functionality. Different computational methods (e.g., protein modelling, substrate-active site docking) supported by experimental data (expression assays) can be extensively used to predict the functional status of specific regions in the genome. Considering this, the expression of NAMPT and nicotinamidase genes (model b) was accessed. Some species had a simultaneously expression of both enzymes and NAMPT protein sequences were extremely conserved, while Nicotinamidases conservation was structural. Species conservation of the catalytic residues of both enzymes was preserved, suggesting that both can be concurrently active. The roles of NAD metabolic enzymes in metabolism or gene expression can be better understood using structural and functional characterization of NAD salvage enzymes. Additionally, as these enzymes are implicated in cancer, diabetes, cardiovascular or neurodegenerative disorders, and also parasitic and infectious diseases, knowledge of their conservation patterns can contribute to targeted drug design. The disruption that occurs in specific segments of a protein with functional relevance (e.g., active site) can also occur in DNA non-B conformations. At a fine-scale, protein systems share some features with non-coding regions showing non-B conformations, including the possibility of mutations that changes the structure of the molecule and specific interactions as H-bonds between atoms. Considering this, the analysis of different interactions that are legible to form non-B DNA conformations (e.g., non-canonical and canonical base-pairing) was analysed at a structural level in two models (model c and d).

12.3. MtDNA Control Region (model c)

Studying the protein-coding regions was of great importance when trying to analyse the non-coding regions. Different mutational signatures can disrupt both proteins and non-B conformations in non-coding regions. The hypothesis to prove was that structured features of coding genome were also present in the non-coding part (sometimes resulting from evolutionary origin out of coding regions, as in model a). The non-coding genome was assumed to present different structural forms (e.g., non-B DNA conformations). The initial step was to develop a method to analyse the putative influence of non-coding regions in coding sequences or important biological processes involved in protein coding mechanisms (e.g., replication, deletions). The focus of the second part of the thesis (model c and d) was the detection of non-B DNA conformations, in the mitochondrial genome [105, 141-144] (model c) and nuclear Y-STRs (model d). The presence of these structures is very difficult to ascertain and there are few accurate experimental procedures [142, 145-147] to elucidate their properties and role in biological processes and only recently the first *in vivo* DNA quadruplex structure was described [148]. Using predictive techniques and algorithms, a full genome annotation of these structures along mitochondrial genome was performed. Some of these structures are formed when mtDNA presents a single-stranded conformation, during replication and transcription. The mechanisms that are associated with already described deletions [94, 149] mainly located in the major arc of the mtDNA between the two proposed origins of replication (OH and OL), results from some specific secondary structures. It was clear that relevant non-B DNA conformations predicted by UNAFold were present both in protein coding and non-coding regions. The detected non-B DNA conformations were clearly associated with different diseases (e.g., chronic progressive external ophthalmoplegia, Kearns–Sayre syndrome or Pearson syndrome, complex multi-system disorder, autosomal disorders). In humans mtDNA control region, the cloverleaf structure at 16071 (Structure A) is supposed to have impact in mtDNA biological processes, what results in deleted molecules [94, 150-156]. Structure A central hairpin is the region of the cloverleaf structure with higher proportion of sites with breakpoints, which means that local specificities of different conformations have impact in the generation of breakpoint deletion patterns. This structural conformation was very stable in molecular dynamics simulations (data not published).

The presence of non-B DNA conformations (e.g., hairpin structures) was not associated only to non-coding regions that are not transcribed, but also to transcribed ones like the WANCY region. This region with tRNA genes presented a free energy variation as high as -36.41 kcal/mol as consequence of formation of several hairpin structures.

The predicted non-B DNA conformations were clearly associated to deletion breakpoints, and this association was demonstrated with statistical significant values

(random distributions versus observed distributions). Structure A [59] and WANCY[144] presented high free energetic folding potential that results in cloverleaf or hairpin structures, and the number deletions breakpoints located at these regions were the highest. Structure A location is outside the important three-stranded D-loop structure (formed in mtDNA replication process), but near enough to infer a putative link between deletion formation and the functional/structural features of the D-loop [157]. The implications of secondary structures can be extended to mechanisms occurring often in mtDNA, like segregation and replication [158]. The impact of the presence of secondary structures resulting from regions with high folding potential in single-stranded states, can have several implications to critical mechanisms in the cell (e.g., replication, transcription), depending on the local conformation adopted (e.g., one hairpin or several hairpins, cloverleaf structures), and in the region were they occur (e.g., functional relevant regions).

12.4. Short Tandem Repeats (model d)

The extension of the non-B conformation analysis to other non-coding regions is further discussed. Small tandemly repeated DNA motifs (model d), known as STRs, in each replication cycle can increase or decrease the number of motifs in each locus although the mechanism is not well understood. Here, this issue was addressed by performing molecular dynamics simulations in tetranucleotide STRs from the Y chromosome. Overall, our results point to the formation of small hairpins (stem base pairing followed by loop nucleation) associated with nearby electronegative pockets of Na+ across the entire sequence of the STR with the exception of the 5' and 3' ends. Our analysis suggests that formation of small hairpins can occur in any region of the STR, and depends on specific conditions of the STR region (base composition, counterions, and water distribution). The process of hairpin formation was associated to electronegative pockets of Na+ that were present near hairpins stems or loops. Different folding potentials will influence STRs length variation resulting from replication errors (e.g., strand-slippage replication mechanism).

Neurodegenerative disorders associated to genomic repetitive segments can be studied using the analysis here performed. The structural dynamics of these genomic regions can be also extended to the study of processes related to other disease-related expansions.

12.4.1. Variation of Free Energy in STRs

Evaluation of the conformational free energy differences along trajectories of each STR model using the MMPBSA was also made as implemented in AmberTools. As the internal dielectric constants is highly dependent of the system we tested three different values (Ɛ=2, 3 and 4). We have verified that the best results were obtained for Ɛ=2 (Table 5).

Table 5: Values of enthalpy variation (ΔH), entropy variation (ΔS), free energy variation (ΔG), and free energy variation relative to single-stranded DNA (ΔΔG) for tested molecular systems, calculated in Amber. *The reference is single-stranded DNA (SS).

Molecular model analysed file	ΔH (kcal/mol)	ΔS (kcal/mol)	ΔG (kcal/mol)	ΔΔG* (kcal/mol)
DYS19A-SS_complex.prmtop	-9304.87	1508.48	-10813.36	0.00
DYS19A-4b13b-5b12b-UF_complex.prmtop	-9267.34	1517.03	-10784.37	29.03
DYS19A-12b21b-13b20b-UFR_complex.prmtop	-9323.93	1520.70	-10844.64	-31.24
DYS19A-20b29b-21b28b-UFR_complex.prmtop	-9054.55	1520.00	-10574.55	238.85
DYS19A-24b33b-25b32b-UFR_complex.prmtop	-9341.26	1524.16	-10865.41	-52.01
DYS19A-28b37b-29b36b-UFR_complex.prmtop	-9234.64	1521.49	-10756.13	57.27
DYS19A-32b41b-33b40b-UFR_complex.prmtop	-9290.29	1519.52	-10809.80	3.60
DYS19A-36b45b-37b44b-UFR_complex.prmtop	-9246.55	1541.13	-10787.68	25.72
DYS19A-40b49b-41b48b-UFR_complex.prmtop	-9209.61	1528.73	-10738.35	75.06
DYS19A-44b53b-45b52b-UFR_complex.prmtop	-7195.25	1529.31	-8724.56	2088.84
DYS19A-48b57b-49b56b-UFR_complex.prmtop	-7771.96	1523.36	-9295.32	1518.08
DYS391TCTA10-SS_complex.prmtop	-6066.25	1010.63	-7076.88	0.00
DYS391TCTA10-4b13b-5b12b-UF_complex.prmtop	-4531.9	1003.22	-5535.12	1541.76
DYS391TCTA10-8b17b-9b16b-UFR_complex.prmtop	-6031.47	1005.53	-7037.00	39.88
DYS391TCTA10-12b21b-13b20b-UFR_complex.prmtop	-5973.14	1004.93	-6978.07	98.81
DYS391TCTA10-16b25b-17b24b-UFR_complex.prmtop	-6047.91	1004.86	-7052.76	24.12
DYS391TCTA10-20b29b-21b28b-UFR_complex.prmtop	-5962.44	1005.28	-6967.72	109.16
DYS391TCTA10-24b33b-25b32b-UFR_complex.prmtop	-5986.24	1015.97	-7002.20	74.68
DYS391TCTA10-28b37b-29b36b-UFR_complex.prmtop	-6041.46	1003.89	-7045.35	31.53
DYS531AAAT11-SS_complex.prmtop	-2358.48	1102.36	-3460.84	0.00
DYS531AAAT11-3b12b-4b11b-UF_complex.prmtop	-6275.11	1127.10	-7402.21	-3941.37
DYS531AAAT11-7b16b-8b15b-UFR_complex.prmtop	-6261.37	1118.29	-7379.66	-3918.82
DYS531AAAT11-11b20b-12b19b-UFR_complex.prmtop	-4454.02	1106.91	-5560.93	-2100.09
DYS531AAAT11-15b24b-16b23b-UFR_complex.prmtop	-6111.17	1120.07	-7231.24	-3770.40
DYS531AAAT11-19b28b-20b27b-UFR_complex.prmtop	-6128.68	1108.11	-7236.79	-3775.95
DYS531AAAT11-23b32b-24b31b-UFR_complex.prmtop	-6225.21	1119.31	-7344.52	-3883.68
DYS531AAAT11-27b36b-28b35b-UFR_complex.prmtop	-4072.23	1097.84	-5170.07	-1709.23
DYS531AAAT11-31b40b-32b39b-UFR_complex.prmtop	-6243.22	1107.22	-7350.44	-3889.60
DYS531AAAT11-35b44b-36b43b-UFR_complex.prmtop	-6245.15	1117.96	-7363.11	-3902.27

The H-bond between base pairs can represent significant variations in energetic values (-2 to -3 kcal/mol/H-bond) [159, 160]. If we consider the total interactions of a base pair the binding energies range from -5 to -47 kcal/mol [159-161]. In the DYS19 we observed variations that represent the different number of H-bonds between the systems and/or hairpin presence (ΔG between 29 kcal/mol to 75 kcal/mol). Nevertheless the DYS19A-44b53b-45b52b-UFR and DYS19A-48b57b-49b56b-UFR complexes presented a higher deviation from the single-stranded DNA molecule (DYS19A-SS) that could not be explained by the change in the number of connected bases and respective H-bonds. These two models presented a super-coiled conformation where coexisting domains of extended and supercoiled DNA are present [162-169]. The DYS391 variation of free energies ranges from 24.12 to 109.16, except for the UNAFold predict model. This model also presented a supercoiled conformation. DYS531 presents a different pattern from the other STRs. When comparing free energies of the models with the single-stranded DNA model, we obtained high free energy differences that were of order ≈4000 kcal/mol for five models and ≈2000 kcal/mol for the others. The high variations in free energy were associated with drastic conformational changes between the different tested models when considering the same STR. The results for each STR demonstrate that the DNA single-stranded states can adopt very different relaxed and supercoiled domains but these energetic states cannot be measure accurately by MMPBSA because these are very different three-dimensional molecular systems. Nevertheless, the theoretical framework behind our computational analysis might be of interest for the analysis of DNA molecular systems and biological processes that can influence DNA mutational patterns.

12.5. SPInDel Workbench

The analysis of specific DNA regions (e.g., non-coding regions) can be performed by different computational programs. Nevertheless, it is usual that a specific analysis of coding regions is performed in a software, and other software used to analyse non-coding regions. This limitation can be overcome and the SPInDel approach was designed using a concept where different regions of the genome and different species can be analysed if they possess a similar evolutionary behaviour. The main purpose was to obtain easy-to-use software with no charges to the user (open source). Using rRNA regions the discrimination of each species was implemented based in the indel variation observed in the alignment. The software can be easily applied to fields like forensics and ecology, since it is optimized for multiple platforms, demonstrates high performance, even with larger databases, and can be modified by the user to perform user custom functions. An extensive database with more than 1,800 species-specific profiles from 18 major taxonomic groups was included in the software, with particular relevance for critically endangered species (e.g., the Bactrian camel, Sumatran orang-utan, kakapo, blue whale, Asian elephant, giant panda, tiger and bonobo). High-throughput analysis of samples in wildlife investigations can be improved using the SPInDel workbench, and the analysed data can be extended to non-coding regions showing indel variation through the species alignments.

13. Concluding Remarks and Future Perspectives

This work started to find shared features between protein-coding (model a and b) and non-coding genomes (model c and d). The putative conversion between the two regions (model a) [170, 171], and the presence of three-dimensional structures of DNA, as the ones occurring in protein systems (model b), was the starting point to address how the different structural DNA molecules behave in different parts of non-coding genome (mtDNA control region and Y-STRs). The relevant non-B conformations can adopt different conformations, as in proteins molecular systems, as demonstrated in this thesis. Although there are specific structural features of proteins (e.g., active site, amino acid residues H-bonds interactions) and of non-B DNA structures (e.g., hairpins, base pairing H-bonds interactions) the two different molecular systems can adopt three-dimensional conformations. In non-coding regions, the formation of non-B conformations has implications in evolution, deletions, replication and disease (model c and d). The role of non-coding structural features in evolution and disease was established for both mtDNA and Y-STRs.

The non-coding structural features were analysed using the knowledge about protein-coding regions molecular systems and the relationships observed between these two types of regions. The phylogeny and evolution of each of these DNA genomic regions demonstrated that the processes modulating the two are not so different. The biological signatures of non-coding regions were detected at the structural conformations observed in DNA sequences. The detected non-B conformations, both in mtDNA and STRs non-coding sequences, were putatively linked to specific processes as replication and deletions. The ENCODE project defined a functional element as a discrete genome segment that encodes a defined product (protein or RNA) or displays a reproducible biochemical signature (e.g., protein binding, or a specific chromatin structure) [20]. Here the definition was extended to specify non-coding DNA regions (e.g., mtDNA control region, STRs) with folding properties that can influence biological processes and by this way present biochemical signatures (non-B DNA conformations), and might even influence gene regulation, replication, and transcription. The specificities of non-coding elements with regulatory properties (e.g., RNAs, transcription factors) that were associated with histone modifications, DNase I hypersensitive sites and DNA methylation, challenged the classical view of non-coding regions. The non-B DNA conformations here detected and analysed, can have impact in biological processes and present a biochemical signature, which means that they can be defined as a functional element, or at least influence processes related with functional elements.

By studying different functional elements of the human genome we have demonstrated that the boundary between coding and non-coding regions is small and depends on a large network of interactions related with different processes (e.g., active site binding, transcription, replication, mutational patterns, tissue specific expression, and, last but not least, the evolutionary time depth of transcriptional loss).The relationships between non-coding and coding genome are extensive and have critical importance to the processes described in Figure 5.

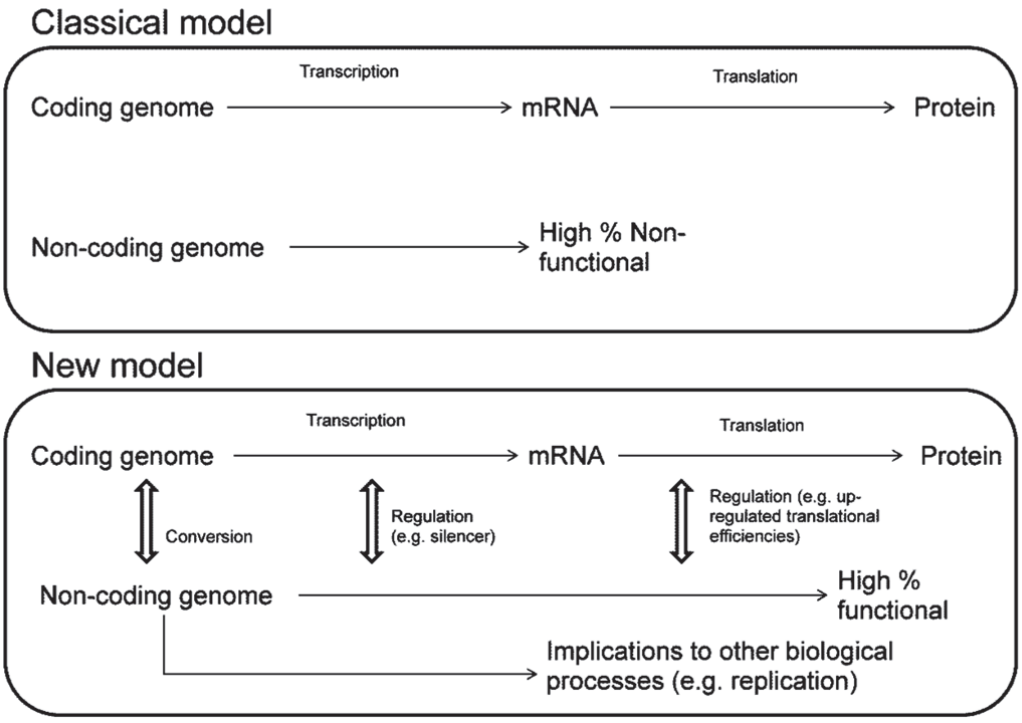


Figure 6: Interactions between coding and non-coding genome.

Non-canonical DNA structures are prone to happen in non-coding regions and are highly relevant to diverse processes occurring in cells as demonstrated in this thesis. Moreover the structures present in each DNA genomic region analysed here can have multiple roles and present a diversity of conformations. Recently guanine-rich regions that are over-represented in telomeres, mitotic and meiotic double-strand break sites, and transcriptional start sites, were characterized *in vitro* and have the potential to form G-quadruplex structure [148, 172-175]. These findings have demonstrated the *in vivo* occurrence of non-canonical DNA structures in different regions of the non-coding genome, usually associated to DNA repetitive motifs. The exhaustive study of these repetitive regions in non-coding regions is of main importance to understand deeply the different mechanisms that regulate biological processes. Our future research will be focused in a global screening

of STR in both nuclear and mitochondrial genome to understand the different conformational changes of non-canonical structures formed in these regions. The topological flexibility of these regions prone to form non-canonical DNA structures can be studied by free energy calculations of the different adopted states at different times. The different conformational changes should be measured by thermodynamic integration or free energy perturbation calculations [176-179] of highly stable molecular systems. These results will be compared with the observed length heterogeneity in each STR.

The highly stable structure A [59, 144] observed in human mitochondrial non-coding genome is also a good starting point to see how specific non-canonical structures behave during critical biological processes. Preliminary results from molecular dynamics simulations with explicit solvent shows the stability of the structure, and simultaneously the flexibility of some stems and loops. To test different stress conditions that result in structure A conformational changes, a steered molecular dynamics (SMD) [180-182] approach can be used. Each loop and stem can be subjected to different forces to ascertain the critical points of this non-canonical structure. On the other hand, the binding properties of structure A can give a new understanding of how specific regions of non-coding mtDNA control region interact with some proteins (e.g., polymerase) or transcription factors. These type of interactions were demonstrated by Eun-Ang Raiber et al. [183] with a non-canonical DNA structure that binds a transcription factor *in vitro*. Using SMD, the full protein-DNA binding landscape can be accessed [184, 185].

14. General Introduction and Discussion References

1. Humphrey, W., A. Dalke, and K. Schulten, *VMD: visual molecular dynamics.* J Mol Graph, 1996. **14**(1): p. 33-8, 27-8.
2. Hsin, J., et al., *Using VMD: an introductory tutorial.* Curr Protoc Bioinformatics, 2008. **Chapter 5**: p. Unit 5 7.
3. Revollo, J.R., A.A. Grimm, and S. Imai, *The regulation of nicotinamide adenine dinucleotide biosynthesis by Nampt/PBEF/visfatin in mammals.* Curr Opin Gastroenterol, 2007. **23**(2): p. 164-70.
4. Watson, J.D. and F.H. Crick, *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.* Nature, 1953. **171**(4356): p. 737-8.
5. Fraser, R.D., *The structure of deoxyribose nucleic acid.* J Struct Biol, 2004. **145**(3): p. 184-5.
6. Alexandrov, B.S., et al., *DNA dynamics is likely to be a factor in the genomic nucleotide repeats expansions related to diseases.* PLoS One, 2011. **6**(5): p. e19800.
7. Alexandrov, B.S., et al., *Bubble statistics and dynamics in double-stranded DNA.* Phys Rev E Stat Nonlin Soft Matter Phys, 2006. **74**(5 Pt 1): p. 050901.
8. Locey, K.J. and E.P. White, *Simple structural differences between coding and noncoding DNA.* PLoS One, 2011. **6**(2): p. e14651.
9. Rogozin, I.B., et al., *Origin and evolution of spliceosomal introns.* Biol Direct, 2012. **7**: p. 11.
10. Parker, S.C., et al., *Local DNA topography correlates with functional noncoding regions of the human genome.* Science, 2009. **324**(5925): p. 389-92.
11. Greenbaum, J.A., B. Pang, and T.D. Tullius, *Construction of a genome-scale structural map at single-nucleotide resolution.* Genome Res, 2007. **17**(6): p. 947-53.
12. Natoli, G. and J.C. Andrau, *Noncoding transcription at enhancers: general principles and functional models.* Annu Rev Genet, 2012. **46**: p. 1-19.
13. Consortium, E.P., *The ENCODE (ENCyclopedia Of DNA Elements) Project.* Science, 2004. **306**(5696): p. 636-40.
14. Consortium, E.P., et al., *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): p. 799-816.
15. de Souza, N., *The ENCODE project.* Nat Methods, 2012. **9**(11): p. 1046.
16. Farnham, P.J., *Thematic minireview series on results from the ENCODE Project: Integrative global analyses of regulatory regions in the human genome.* J Biol Chem, 2012. **287**(37): p. 30885-7.
17. Harrow, J., et al., *GENCODE: the reference human genome annotation for The ENCODE Project.* Genome Res, 2012. **22**(9): p. 1760-74.
18. Pennisi, E., *Genomics. ENCODE project writes eulogy for junk DNA.* Science, 2012. **337**(6099): p. 1159, 1161.
19. Sastre, L., *Clinical implications of the ENCODE project.* Clin Transl Oncol, 2012. **14**(11): p. 801-2.
20. Consortium, E.P., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
21. Orgel, L.E. and F.H. Crick, *Selfish DNA: the ultimate parasite.* Nature, 1980. **284**(5757): p. 604-7.
22. Cavalier-Smith, T., *Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox.* J Cell Sci, 1978. **34**: p. 247-78.

23. Lynch, M. and J.S. Conery, *The origins of genome complexity.* Science, 2003. **302**(5649): p. 1401-4.
24. Janicki, M., R. Rooke, and G. Yang, *Bioinformatics and genomic analysis of transposable elements in eukaryotic genomes.* Chromosome Res, 2011. **19**(6): p. 787-808.
25. Redi, C.A. and E. Capanna, *Genome size evolution: sizing mammalian genomes.* Cytogenet Genome Res, 2012. **137**(2-4): p. 97-112.
26. Dufresne, F. and N. Jeffery, *A guided tour of large genome size in animals: what we know and where we are heading.* Chromosome Res, 2011. **19**(7): p. 925-38.
27. Hillier, L.W., et al., *Genomics in C. elegans: so many genes, such a little worm.* Genome Res, 2005. **15**(12): p. 1651-60.
28. Hughes, A.L., *Adaptive evolution after gene duplication.* Trends Genet, 2002. **18**(9): p. 433-4.
29. Lynch, M. and V. Katju, *The altered evolutionary trajectories of gene duplicates.* Trends Genet, 2004. **20**(11): p. 544-9.
30. Ohta, T., *Simulating evolution by gene duplication.* Genetics, 1987. **115**(1): p. 207-13.
31. Prince, V.E. and F.B. Pickett, *Splitting pairs: the diverging fates of duplicated genes.* Nat Rev Genet, 2002. **3**(11): p. 827-37.
32. Force, A., et al., *Preservation of duplicate genes by complementary, degenerative mutations.* Genetics, 1999. **151**(4): p. 1531-45.
33. Lynch, M., et al., *The probability of preservation of a newly arisen gene duplicate.* Genetics, 2001. **159**(4): p. 1789-804.
34. Lynch, M. and A. Force, *The probability of duplicate gene preservation by subfunctionalization.* Genetics, 2000. **154**(1): p. 459-73.
35. Maroni, G., et al., *Metallothionein gene duplications and metal tolerance in natural populations of Drosophila melanogaster.* Genetics, 1987. **117**(4): p. 739-44.
36. Egli, D., et al., *The four members of the Drosophila metallothionein family exhibit distinct yet overlapping roles in heavy metal homeostasis and detoxification.* Genes Cells, 2006. **11**(6): p. 647-58.
37. Revollo, J.R., A.A. Grimm, and S. Imai, *The NAD biosynthesis pathway mediated by nicotinamide phosphoribosyltransferase regulates Sir2 activity in mammalian cells.* J Biol Chem, 2004. **279**(49): p. 50754-63.
38. Hara, N., et al., *Elevation of cellular NAD levels by nicotinic acid and involvement of nicotinic acid phosphoribosyltransferase in human cells.* J Biol Chem, 2007. **282**(34): p. 24574-82.
39. Belenky, P., et al., *Nicotinamide riboside and nicotinic acid riboside salvage in fungi and mammals. Quantitative basis for Urh1 and purine nucleoside phosphorylase function in NAD+ metabolism.* J Biol Chem, 2009. **284**(1): p. 158-64.
40. Lin, H., A.L. Kwan, and S.K. Dutcher, *Synthesizing and salvaging NAD: lessons learned from Chlamydomonas reinhardtii.* PLoS Genet, 2010. **6**(9).
41. Sorci, L., et al., *Genomics-driven reconstruction of acinetobacter NAD metabolism: insights for antibacterial target selection.* J Biol Chem, 2010. **285**(50): p. 39490-9.
42. Michels, P.A. and L. Avilan, *The NAD+ metabolism of Leishmania, notably the enzyme nicotinamidase involved in NAD+ salvage, offers prospects for development of anti-parasite chemotherapy.* Mol Microbiol, 2011. **82**(1): p. 4-8.
43. Gazanion, E., et al., *The Leishmania nicotinamidase is essential for NAD+ production and parasite proliferation.* Mol Microbiol, 2011. **82**(1): p. 21-38.
44. Jewett, M.W., et al., *Molecular characterization of the Borrelia burgdorferi in vivo-essential protein PncA.* Microbiology, 2011. **157**(Pt 10): p. 2831-40.
45. Domergue, R., et al., *Nicotinic acid limitation regulates silencing of Candida adhesins during UTI.* Science, 2005. **308**(5723): p. 866-70.
46. Seiner, D.R., S.S. Hegde, and J.S. Blanchard, *Kinetics and inhibition of nicotinamidase from Mycobacterium tuberculosis.* Biochemistry, 2010. **49**(44): p. 9613-9.
47. Zhang, H., et al., *Characterization of Mycobacterium tuberculosis nicotinamidase/pyrazinamidase.* FEBS J, 2008. **275**(4): p. 753-62.

48. Gallo, C.M., D.L. Smith, Jr., and J.S. Smith, *Nicotinamide clearance by Pnc1 directly regulates Sir2-mediated silencing and longevity.* Mol Cell Biol, 2004. **24**(3): p. 1301-12.
49. Balan, V., et al., *Life span extension and neuronal cell protection by Drosophila nicotinamidase.* J Biol Chem, 2008. **283**(41): p. 27810-9.
50. Mesko, B., et al., *Peripheral blood gene expression patterns discriminate among chronic inflammatory diseases and healthy controls and identify novel targets.* BMC Med Genomics, 2010. **3**: p. 15.
51. Galli, M., et al., *The nicotinamide phosphoribosyltransferase: a molecular link between metabolism, inflammation, and cancer.* Cancer Res, 2010. **70**(1): p. 8-11.
52. Silva, R.M., et al., *The yeast PNC1 longevity gene is up-regulated by mRNA mistranslation.* PLoS One, 2009. **4**(4): p. e5212.
53. Burnett, C., et al., *Absence of effects of Sir2 overexpression on lifespan in C. elegans and Drosophila.* Nature, 2011. **477**(7365): p. 482-5.
54. Chinnery, P., *Mitochondrial DNA in Homo Sapiens*, in *Human Mitochondrial DNA and the Evolution of Homo sapiens*. 2006, Springer-Verlag Berlin Heidelberg: Germany.
55. Jobling, M.A., M.E. Hurles, and C. Tyler-Smith, *Human Evolutionary Genetics: Origins, Peoples & Disease*. Human Evolutionary Genetics: Origins, Peoples & Disease. 2004: Garland Science. 39-43.
56. Case, J. and D. Wallace, *Maternal inheritance of mitochondrial DNA polymorphisms in cultured human fibroblasts.* Somatic Cell Genet., 1981(Jan;7(1)): p. 103-8.
57. Giles, R., et al., *Maternal inheritance of human mitochondrial DNA.* Proc Natl Acad Sci U S A, 1980. **77**(11): p. 6715-9.
58. Hutchison, C.r., et al., *Maternal inheritance of mammalian mitochondrial DNA.* Nature, 1974. **251**(5475): p. 536-8.
59. Pereira, F., et al., *Evidence for variable selective pressures at a large secondary structure of the human mitochondrial DNA control region.* Mol Biol Evol, 2008. **25**(12): p. 2759-70.
60. Fan, H. and J.Y. Chu, *A brief review of short tandem repeat mutation.* Genomics Proteomics Bioinformatics, 2007. **5**(1): p. 7-14.
61. Farre, M., et al., *Assessing the role of tandem repeats in shaping the genomic architecture of great apes.* PLoS One, 2011. **6**(11): p. e27239.
62. Kayser, M. and A. Sajantila, *Mutations at Y-STR loci: implications for paternity testing and forensic analysis.* Forensic Sci Int, 2001. **118**(2-3): p. 116-21.
63. Pereira, L., M.J. Prata, and A. Amorim, *Mismatch distribution analysis of Y-STR haplotypes as a tool for the evaluation of identity-by-state proportions and significance of matches--the European picture.* Forensic Sci Int, 2002. **130**(2-3): p. 147-55.
64. Butler, J.M., et al., *Chromosomal duplications along the Y-chromosome and their potential impact on Y-STR interpretation.* J Forensic Sci, 2005. **50**(4): p. 853-9.
65. Decker, A.E., et al., *Analysis of mutations in father-son pairs with 17 Y-STR loci.* Forensic Sci Int Genet, 2008. **2**(3): p. e31-5.
66. Balaresque, P., et al., *Genomic complexity of the Y-STR DYS19: inversions, deletions and founder lineages carrying duplications.* Int J Legal Med, 2009. **123**(1): p. 15-23.
67. Shi, W., et al., *A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations.* Mol Biol Evol, 2010. **27**(2): p. 385-93.
68. Ruitberg, C.M., D.J. Reeder, and J.M. Butler, *STRBase: a short tandem repeat DNA database for the human identity testing community.* Nucleic Acids Res, 2001. **29**(1): p. 320-2.
69. Jobling, M.A., *Human Evolutionary Genetics: Origins, Peoples & Disease*, ed. M.A.e.a. Jobling. 2004, New York, USA.: Garland Science.
70. Hubscher, U., G. Maga, and S. Spadari, *Eukaryotic DNA polymerases.* Annu Rev Biochem, 2002. **71**: p. 133-63.
71. Sakaguchi, K., F. Sugawara, and Y. Mizushina, *[Inhibitors of eukaryotic DNA polymerases].* Seikagaku, 2002. **74**(3): p. 244-51.
72. Caliebe, A., et al., *A Markov chain description of the stepwise mutation model: local and global behaviour of the allele process.* J Theor Biol, 2010. **266**(2): p. 336-42.

73. Fu, Y.X. and R. Chakraborty, *Simultaneous estimation of all the parameters of a stepwise mutation model.* Genetics, 1998. **150**(1): p. 487-97.
74. Fuerst, P.A. and R.E. Ferrell, *The stepwise mutation model: an experimental evaluation utilizing hemoglobin variants.* Genetics, 1980. **94**(1): p. 185-201.
75. Gusmao, L., et al., *Bimodal allele frequency distribution at Y-STR loci DYS392 and DYS438: no evidence for a deviation from the stepwise mutation model.* Int J Legal Med, 2003. **117**(5): p. 287-90.
76. Kimmel, M. and R. Chakraborty, *Measures of variation at DNA repeat loci under a general stepwise mutation model.* Theor Popul Biol, 1996. **50**(3): p. 345-67.
77. Kimura, M. and T. Ohta, *Stepwise mutation model and distribution of allelic frequencies in a finite population.* Proc Natl Acad Sci U S A, 1978. **75**(6): p. 2868-72.
78. Valdes, A.M., M. Slatkin, and N.B. Freimer, *Allele frequencies at microsatellite loci: the stepwise mutation model revisited.* Genetics, 1993. **133**(3): p. 737-49.
79. Canceill, D., E. Viguera, and S.D. Ehrlich, *Replication slippage of different DNA polymerases is inversely related to their strand displacement efficiency.* J Biol Chem, 1999. **274**(39): p. 27481-90.
80. MacLean, H.E., et al., *Double-strand DNA break repair with replication slippage on two strands: a novel mechanism of deletion formation.* Hum Mutat, 2006. **27**(5): p. 483-9.
81. Boyer, J.C., et al., *Fidelity of DNA replication by extracts of normal and malignantly transformed human cells.* Cancer Res, 1993. **53**(14): p. 3270-5.
82. Kunkel, T.A., et al., *Analysis of fidelity mechanisms with eukaryotic DNA replication and repair proteins.* Genome, 1989. **31**(1): p. 100-3.
83. Thomas, D.C., et al., *Fidelity of animal cell DNA polymerases alpha and delta and of a human DNA replication complex.* Basic Life Sci, 1990. **52**: p. 289-97.
84. Thomas, D.C., et al., *Fidelity of mammalian DNA replication and replicative DNA polymerases.* Biochemistry, 1991. **30**(51): p. 11751-9.
85. Pereira, F., J. Carneiro, and A. Amorim, *Identification of species with DNA-based technology: current progress and challenges.* Recent Pat DNA Gene Seq, 2008. **2**(3): p. 187-99.
86. Bacolla, A. and R.D. Wells, *Non-B DNA conformations as determinants of mutagenesis and human disease.* Mol Carcinog, 2009. **48**(4): p. 273-85.
87. Wells, R.D., *Non-B DNA conformations, mutagenesis and disease.* Trends Biochem Sci, 2007. **32**(6): p. 271-8.
88. Wright, B.E., *A biochemical mechanism for nonrandom mutations and evolution.* J Bacteriol, 2000. **182**(11): p. 2993-3001.
89. Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic Acids Res, 2003. **31**(13): p. 3406-15.
90. Rannala, B. and Z. Yang, *Phylogenetic inference using whole genomes.* Annu Rev Genomics Hum Genet, 2008. **9**: p. 217-31.
91. Takezaki, N. and M. Nei, *Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA.* Genetics, 1996. **144**(1): p. 389-99.
92. Yang, Z., N. Goldman, and A. Friday, *Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation.* Mol Biol Evol, 1994. **11**(2): p. 316-24.
93. Pereira, F., et al., *Evidence for Variable Selective Pressures at a Large Secondary Structure of the Human Mitochondrial DNA Control Region.* Molecular Biology and Evolution, 2008. **25**(12): p. 2759-2770.
94. Samuels, D.C., E.A. Schon, and P.F. Chinnery, *Two direct repeats cause most human mtDNA deletions.* Trends Genet, 2004. **20**(9): p. 393-8.
95. Bacolla, A., et al., *Breakpoints of gross deletions coincide with non-B DNA conformations.* Proc Natl Acad Sci U S A, 2004. **101**(39): p. 14162-7.
96. Bacolla, A. and R.D. Wells, *Non-B DNA conformations, genomic rearrangements, and human disease.* J Biol Chem, 2004. **279**(46): p. 47411-4.
97. Krishnan, K.J., et al., *What causes mitochondrial DNA deletions in human cells?* Nat Genet, 2008. **40**(3): p. 275-9.

98. Pavletich, N.P. and C.O. Pabo, *Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers.* Science, 1993. **261**(5129): p. 1701-7.
99. Rohs, R., et al., *Nuance in the double-helix and its role in protein-DNA recognition.* Curr Opin Struct Biol, 2009. **19**(2): p. 171-7.
100. Voineagu, I., et al., *Replication stalling at unstable inverted repeats: interplay between DNA hairpins and fork stabilizing proteins.* Proc Natl Acad Sci U S A, 2008. **105**(29): p. 9936-41.
101. Nelson, J.W., F.H. Martin, and I. Tinoco, Jr., *DNA and RNA oligomer thermodynamics: the effect of mismatched bases on double-helix stability.* Biopolymers, 1981. **20**(12): p. 2509-31.
102. Ratilainen, T. and B. Norden, *Thermodynamics of PNA interactions with DNA and RNA.* Methods Mol Biol, 2002. **208**: p. 59-88.
103. Sugimoto, N., M. Nakano, and S. Nakano, *Thermodynamics-structure relationship of single mismatches in RNA/DNA duplexes.* Biochemistry, 2000. **39**(37): p. 11270-81.
104. Xia, T., J.A. McDowell, and D.H. Turner, *Thermodynamics of nonsymmetric tandem mismatches adjacent to G.C base pairs in RNA.* Biochemistry, 1997. **36**(41): p. 12486-97.
105. Markham, N.R. and M. Zuker, *UNAFold: software for nucleic acid folding and hybridization.* Methods Mol Biol, 2008. **453**: p. 3-31.
106. SantaLucia, J., Jr. and D. Hicks, *The thermodynamics of DNA structural motifs.* Annu Rev Biophys Biomol Struct, 2004. **33**: p. 415-40.
107. Meroueh, M. and C.S. Chow, *Thermodynamics of RNA hairpins containing single internal mismatches.* Nucleic Acids Res, 1999. **27**(4): p. 1118-25.
108. Mascotti, D.P. and T.M. Lohman, *Thermodynamics of oligoarginines binding to RNA and DNA.* Biochemistry, 1997. **36**(23): p. 7272-9.
109. SantaLucia, J., Jr. and D.H. Turner, *Measuring the thermodynamics of RNA secondary structure formation.* Biopolymers, 1997. **44**(3): p. 309-19.
110. Qiu, H., et al., *Thermodynamics of folding of the RNA pseudoknot of the T4 gene 32 autoregulatory messenger RNA.* Biochemistry, 1996. **35**(13): p. 4176-86.
111. Case, D.A., et al., *The Amber biomolecular simulation programs.* J Comput Chem, 2005. **26**(16): p. 1668-88.
112. D.A. Case, T.A.D., T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A.W. Goetz, I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, R.M. Wolf, J. Liu, X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P.A. Kollman, *AMBER 12*. 2012: University of California, San Francisco.
113. Tom Macke, W.A.S.-S., Russell A. Brown, Istvan Kolossvary, Yannick Bomble and David A. Case, *Generation of Models for "Unusual" DNA and RNA: A Computer Language for Structural Exploration*.
114. Case, T.M.a.D.A., *Modeling unusual nucleic acid structures*, in *Molecular Modeling of Nucleic Acids*, J. N.B. Leontes and J. SantaLucia, eds. , Editor. 1998, Washington, DC: American Chemical Society. p. pp. 379-393.
115. Brice, A.R. and B.N. Dominy, *Analyzing the robustness of the MM/PBSA free energy calculation method: application to DNA conformational transitions.* J Comput Chem, 2011. **32**(7): p. 1431-40.
116. Warwicker, J. and H.C. Watson, *Calculation of the electric potential in the active site cleft due to alpha-helix dipoles.* J Mol Biol, 1982. **157**(4): p. 671-9.
117. Bashford, D. and D.A. Case, *Generalized born models of macromolecular solvation effects.* Annu Rev Phys Chem, 2000. **51**: p. 129-52.
118. Kollman, P.A., et al., *Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models.* Accounts of Chemical Research, 2000. **33**(12): p. 889-897.
119. Connolly, M.L., *Solvent-accessible surfaces of proteins and nucleic acids.* Science, 1983. **221**(4612): p. 709-13.

120. Excoffier, L. and G. Heckel, *Computer programs for population genetics data analysis: a survival guide.* Nat Rev Genet, 2006. **7**(10): p. 745-58.
121. Kumar, S. and J. Dudley, *Bioinformatics software for biologists in the genomics era.* Bioinformatics, 2007. **23**(14): p. 1713-7.
122. Rappin, N. and R. Dunn, *wxPython in Action*. 2006. 583.
123. Felsenstein, J., *PHYLIP*. 2008, Distributed by the author: Department of Genetics, University of Washington, Seattle. p. Phylogeny Inference Package.
124. Tamura, K., et al., *MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0.* Mol Biol Evol, 2007. **24**(8): p. 1596-9.
125. Yang, Z., *PAML 4: phylogenetic analysis by maximum likelihood.* Mol Biol Evol, 2007. **24**(8): p. 1586-91.
126. Ronquist, F. and J. Huelsenbeck, *MrBayes 3: Bayesian phylogenetic inference under mixed models.* Bioinformatics, 2003. **19**(12): p. 1572-4.
127. Hall, T., *BioEdit*. 2007.
128. Rossum, G.v., *Python Documentation*. Release 2.5.2 ed. Python Documentation, ed. F.L. Drake. 2008: Python Software Foundation.
129. Chang, J., et al., *Biopython Tutorial and Cookbook*. 2008.
130. Knight, R., et al., *PyCogent: a toolkit for making sense from sequence.* Genome Biol, 2007. **8**(8): p. R171.
131. Pritchard, L., et al., *GenomeDiagram: a python package for the visualization of large-scale genomic data.* Bioinformatics, 2006. **22**(5): p. 616-7.
132. Hetland, M.L., *Beginning Python: From Novice to Professional*. Second Edition ed, ed. F. Pohlmann. Vol. 1. 2008: Apress. 656.
133. Bassi, S., *A primer on python for life science researchers.* PLoS Comput Biol, 2007. **3**(11): p. e199.
134. Lavery, R., et al., *Conformational analysis of nucleic acids revisited: Curves+.* Nucleic Acids Res, 2009. **37**(17): p. 5917-29.
135. Lu, X.J. and W.K. Olson, *3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures.* Nucleic Acids Res, 2003. **31**(17): p. 5108-21.
136. Lu, X.J. and W.K. Olson, *3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures.* Nat Protoc, 2008. **3**(7): p. 1213-27.
137. Zheng, G., X.J. Lu, and W.K. Olson, *Web 3DNA--a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W240-6.
138. Li, L., et al., *DelPhi: a comprehensive suite for DelPhi software and associated resources.* BMC Biophys, 2012. **5**(1): p. 9.
139. Chatterjee, S. and J.K. Pal, *Role of 5'- and 3'-untranslated regions of mRNAs in human diseases.* Biol Cell, 2009. **101**(5): p. 251-62.
140. Cawley, A. and J. Warwicker, *eIF4E-binding protein regulation of mRNAs with differential 5'-UTR secondary structure: a polyelectrostatic model for a component of protein-mRNA interactions.* Nucleic Acids Res, 2012. **40**(16): p. 7666-75.
141. Frank-Kamenetskii, M.D. and S.M. Mirkin, *Triplex DNA structures.* Annu Rev Biochem, 1995. **64**: p. 65-95.
142. Mirkin, S.M., *DNA structures, repeat expansions and human hereditary disorders.* Curr Opin Struct Biol, 2006. **16**(3): p. 351-8.
143. Voineagu, I., C.H. Freudenreich, and S.M. Mirkin, *Checkpoint responses to unusual structures formed by DNA repeats.* Mol Carcinog, 2009. **48**(4): p. 309-18.
144. Damas, J., et al., *Mitochondrial DNA deletions are associated with non-B DNA conformations.* Nucleic Acids Res, 2012.
145. Duzdevich, D., et al., *Unusual structures are present in DNA fragments containing super-long Huntingtin CAG repeats.* PLoS One, 2011. **6**(2): p. e17119.

146. Mirkin, S.M., *Expandable DNA repeats and human disease.* Nature, 2007. **447**(7147): p. 932-40.
147. Mirkin, S.M., *DNA Topology: Fundamentals* in *ENCYCLOPEDIA OF LIFE SCIENCES*. 2001, Nature Publishing Group: University of Illinois at Chicago, Illinois, USA.
148. Biffi, G., et al., *Quantitative visualization of DNA G-quadruplex structures in human cells.* Nat Chem, 2013. **advance online publication**.
149. Kajander, O.A., et al., *Human mtDNA sublimons resemble rearranged mitochondrial genoms found in pathological states.* Hum Mol Genet, 2000. **9**(19): p. 2821-35.
150. Lee, Y.S., W.D. Kennedy, and Y.W. Yin, *Structural insight into processive human mitochondrial DNA synthesis and disease-related polymerase mutations.* Cell, 2009. **139**(2): p. 312-24.
151. Tuppen, H.A., et al., *Mitochondrial DNA mutations and human disease.* Biochim Biophys Acta, 2010. **1797**(2): p. 113-28.
152. Greaves, L.C. and R.W. Taylor, *Mitochondrial DNA mutations in human disease.* IUBMB Life, 2006. **58**(3): p. 143-51.
153. Taylor, R.W. and D.M. Turnbull, *Mitochondrial DNA mutations in human disease.* Nat Rev Genet, 2005. **6**(5): p. 389-402.
154. Chinnery, P.F. and D.M. Turnbull, *Mitochondrial DNA mutations in the pathogenesis of human disease.* Mol Med Today, 2000. **6**(11): p. 425-32.
155. Inagaki, H., et al., *Chromosomal instability mediated by non-B DNA: cruciform conformation and not DNA sequence is responsible for recurrent translocation in humans.* Genome Res, 2009. **19**(2): p. 191-8.
156. Hou, J.H. and Y.H. Wei, *The unusual structures of the hot-regions flanking large-scale deletions in human mitochondrial DNA.* Biochem J, 1996. **318 (Pt 3)**: p. 1065-70.
157. Callen, J.C., et al., *Changes in D-loop frequency and superhelicity among the mitochondrial DNA molecules in relation to organelle biogenesis in oocytes of Xenopus laevis.* Exp Cell Res, 1983. **143**(1): p. 115-25.
158. Holt, I.J., et al., *Mammalian mitochondrial nucleoids: organizing an independently minded genome.* Mitochondrion, 2007. **7**(5): p. 311-21.
159. Jurecka, P. and P. Hobza, *True stabilization energies for the optimal planar hydrogen-bonded and stacked structures of guanine...cytosine, adenine...thymine, and their 9- and 1-methyl derivatives: complete basis set calculations at the MP2 and CCSD(T) levels and comparison with experiment.* J Am Chem Soc, 2003. **125**(50): p. 15608-13.
160. Sponer, J., P. Jurecka, and P. Hobza, *Accurate interaction energies of hydrogen-bonded nucleic acid base pairs.* J Am Chem Soc, 2004. **126**(32): p. 10142-51.
161. Kabelac, M., et al., *Structure, energetics, vibrational frequencies and charge transfer of base pairs, nucleoside pairs, nucleotide pairs and B-DNA pairs of trinucleotides: ab initio HF/MINI-1 and empirical force field study.* J Biomol Struct Dyn, 2000. **17**(6): p. 1077-86.
162. Marko, J.F., *Twist and shout (and pull): molecular chiropractors undo DNA.* Proc Natl Acad Sci U S A, 1997. **94**(22): p. 11770-2.
163. Marko, J.F., *Torque and dynamics of linking number relaxation in stretched supercoiled DNA.* Phys Rev E Stat Nonlin Soft Matter Phys, 2007. **76**(2 Pt 1): p. 021926.
164. Marko, J.F. and M.G. Poirier, *Micromechanics of chromatin and chromosomes.* Biochem Cell Biol, 2003. **81**(3): p. 209-20.
165. Marko, J.F. and E.D. Siggia, *Fluctuations and supercoiling of DNA.* Science, 1994. **265**(5171): p. 506-8.
166. Marko, J.F. and E.D. Siggia, *Statistical mechanics of supercoiled DNA.* Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics, 1995. **52**(3): p. 2912-2938.
167. Yan, J., R. Kawamura, and J.F. Marko, *Statistics of loop formation along double helix DNAs.* Phys Rev E Stat Nonlin Soft Matter Phys, 2005. **71**(6 Pt 1): p. 061905.
168. Yan, J., M.O. Magnasco, and J.F. Marko, *Kinetic proofreading can explain the supression of supercoiling of circular DNA molecules by type-II topoisomerases.* Phys Rev E Stat Nonlin Soft Matter Phys, 2001. **63**(3 Pt 1): p. 031909.

169. Yan, J. and J.F. Marko, *Localized single-stranded bubble mechanism for cyclization of short double helix DNA.* Phys Rev Lett, 2004. **93**(10): p. 108108.
170. Moleirinho, A., et al., *Gains, losses and changes of function after gene duplication: study of the metallothionein family.* PLoS One, 2011. **6**(4): p. e18487.
171. Levine, M.T., et al., *Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression.* Proc Natl Acad Sci U S A, 2006. **103**(26): p. 9935-9.
172. Hershman, S.G., et al., *Genomic distribution and functional analyses of potential G-quadruplex-forming sequences in Saccharomyces cerevisiae.* Nucleic Acids Res, 2008. **36**(1): p. 144-56.
173. Huppert, J.L. and S. Balasubramanian, *Prevalence of quadruplexes in the human genome.* Nucleic Acids Res, 2005. **33**(9): p. 2908-16.
174. Huppert, J.L. and S. Balasubramanian, *G-quadruplexes in promoters throughout the human genome.* Nucleic Acids Res, 2007. **35**(2): p. 406-13.
175. Eddy, J. and N. Maizels, *Gene function correlates with potential for G4 DNA formation in the human genome.* Nucleic Acids Res, 2006. **34**(14): p. 3887-96.
176. Guimaraes, C.R. and A.M. Mathiowetz, *Addressing limitations with the MM-GB/SA scoring procedure using the WaterMap method and free energy perturbation calculations.* J Chem Inf Model, 2010. **50**(4): p. 547-59.
177. Pearlman, D.A. and P.S. Charifson, *Are free energy calculations useful in practice? A comparison with rapid scoring functions for the p38 MAP kinase protein system.* J Med Chem, 2001. **44**(21): p. 3417-23.
178. Miyamoto, S. and P.A. Kollman, *Absolute and relative binding free energy calculations of the interaction of biotin and its analogs with streptavidin using molecular dynamics/free energy perturbation approaches.* Proteins, 1993. **16**(3): p. 226-45.
179. Moreira, I.S., P.A. Fernandes, and M.J. Ramos, *Computational alanine scanning mutagenesis--an improved methodological approach.* J Comput Chem, 2007. **28**(3): p. 644-54.
180. Isralewitz, B., et al., *Steered molecular dynamics investigations of protein function.* J Mol Graph Model, 2001. **19**(1): p. 13-25.
181. Isralewitz, B., M. Gao, and K. Schulten, *Steered molecular dynamics and mechanical functions of proteins.* Curr Opin Struct Biol, 2001. **11**(2): p. 224-30.
182. Mascayano, C., et al., *Binding of arachidonic acid and two flavonoid inhibitors to human 12- and 15-lipoxygenases: a steered molecular dynamics study.* J Mol Model, 2010. **16**(5): p. 1039-45.
183. Raiber, E.A., et al., *A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro.* Nucleic Acids Res, 2012. **40**(4): p. 1499-508.
184. Chen, C. and B.M. Pettitt, *The binding process of a nonspecific enzyme with DNA.* Biophys J, 2011. **101**(5): p. 1139-47.
185. Chen, L.Y., *Exploring the free-energy landscapes of biological systems with steered molecular dynamics.* Phys Chem Chem Phys, 2011. **13**(13): p. 6176-83.