UNIVERSITY of FLORENCE
CERM - Magnetic Resonance Center

International Doctorate in Structural Biology

Cycle XX
2005 - 2007

Evaluation and improvement of data quality and methods in structural genomics

Ph.D. thesis of

Edoardo Saccenti

Tutor	Coordinator
Prof. Ivano Bertini	Prof. Claudio Luchinat

S.S.D CHIM/03

This thesis has been approved by the University of Florence,
the University of Frankfurt and the Utrecht University

Florence, January MMVIII

Ancora una volta ai miei genitori

E porterem con rivelazion somma
saper che all'altro somma.
Corrado R. L. dei Toscani

Trahit sua quemque voluptas.
Virgilius

Raffiniert ist der Herr Gott, aber boshaft ist er nicht.
Anonimo

Monsieur, $\frac{1}{n}(a+b^n)=x$, *donc Dieu existe; repondez!*
Leonard Euler to Denis Diderot

Contents

The modern structural biologist

We may, I believe, anticipate that the chemist of the future who is interested in the structure of proteins, nucleic acids, polysaccharides, and other complex substances with high molecular weight will come to rely upon a new structural chemistry, involving precise geometrical relationships among the atoms in the molecules and the rigorous application of the new structural principles, and that great progress will be made, through this technique, in the attack, by chemical methods, on the problems of biology and medicine. With these words, Linus Pauling concluded his Nobel Lecture on December 11, 1954.

A deeper understanding of basic aspects of structure-function relationships in biologically active macromolecules requires that many different yet complimentary approaches be used in modern structural biology. Bioinformatics, genetics and structural determination are nowadays integrated to fully understand the behavior of macromolecules of particular interests such as proteins involved in copper trafficking or in cellular respiration.

With Linus Pauling's caveat in mind, according to whom progresses in biology and medicine will be accomplished only by mean of a full understanding of the physics and chemistry of biological processes, it is clear that a modern structural biologist must encompass skills and knowledge in different yet not related scientific fields. With the completion of Human Genome in 2003, the post-genomic era begun, and for the first time scientists are given all the instruments to understand how Life works. An incredible amount of data is now easily accessible through biological databases. The structural biologist must know how to deal with countless data and this can be only accomplished by mastering bioinformatic tools, not only the existing ones but even developing new tools for both personals and general purposes thus using and integrating, trough an adequate programming language, mathematics to reduce the reality to formulas, biostatistics to extract meaningful and unique information from redundancy and chemistry and biology to interpret his results.

Despite the fact that *in silico* studies could be indicative to trace the perspectives and suggest theoretical approaches for a biological problem, a structural biologist must be able, of course, to perform direct experiments in the laboratory because experiments are the corner stone of the Scientific Method. As structural biology is focused on understanding how proteins works by means of the study of their interactions and their structures, this interest has stimulated the development of a number of biochemical and genetic approaches to identify and clone genes encoding interacting proteins to apply techniques such as co-immunoprecipitation, co-purification and the yeast two-hybrid system, devised to identify genes encoding proteins that physically associate with a given protein *in vivo*.

Once a system and/or its interaction partners has been biochemically and biologically characterized, the knowledge of their structures and the details of their mutual interplay could be assayed using structural methods such as nucler magnetic resonance (NMR) spectroscopy.

As NMR has the unique feature of allowing *in vivo* studies, a structural biologist must be able to understand the fundamentals and the development of the use of NMR spectroscopy for the determination of three-dimensional structures of biological macromolecules in solution. He must not only know the different aspects of the many kinds of experiments that can be performed to investigate a particular physical effect but also the physics of the phenomena he is looking at.

As long as structural biologists rely on experimentally determined three-dimensional structures to link biomolecular data and structural information, he cannot be unaware of the quality of data he derives from public databases. Assessment of NMR structures quality is indeed a long-standing issue; the community of bio-NMR scientists that are focused on protein structural biology is still lacking a generally accepted measure of accuracy of the structures produced, although several measures have been proposed over the years. To have a wide spread scientific insight in different aspect of structural biology, during my three years Ph.D. course, I worked on different projects, focused on bioinformatics, genetic approaches to protein-protein interactions and validation of three dimensional NMR-solved protein structures.

. . . non vogliate negar l'esperïenza
di retro al sol, del mondo sanza gente.
Considerate la vostra semenza
fatti non foste a viver come bruti
ma per seguir virtute e canoscenza

Dante Alighieri, *Divina Commedia*
Inferno canto XXVI, vv. 116-120

Introduction

Computational biology is a highly interdisciplinary field of biology, relying on basic principles from computer science, biology, physics, chemistry, mathematics, and statistics.

Bioinformatics, a popular term in this era of large-scale DNA sequencing, is only a subset of computational biology - the part concerned with the storage, organization, curation and annotation of biological data but is now often used as an umbrella term for almost all aspects of computational biology [66].

Computational biology and bioinformatics originated as a cross-disciplinary field as the need for computational solutions to research problems raised in biology and medicine [112]. These fields evolved as computation became cheaper and widespread during the 80s, as the Internet grew during the 90s, and as high-throughput technologies become common in the 2000s. The boundaries between numerical and biological disciplines have become blurred [108] and indeed, recent years have seen the spawning of bioinformatics sub-disciplines such as cheminformatics [99], neuroinformatics [7] and immunoinformatics [25]. Ostensibly, the only limit to the number of bioinformatics/computational-related sub-disciplines is the number of disciplines themselves. Because of the growing need for an integrative view of biological problems, cross-disciplinary efforts such as these are considered increasingly important to continued scientific progress [130].

Computational biology extends beyond bioinformatics into the realm of sequence analysis: finding genes and ascertaining their function; predicting the structure of proteins and RNA sequences; and determining the evolutionary relationship of proteins and DNA sequences.

Trough computational biology, the biological functions of cells, tissues, and organisms can be explained in terms of the structure and behavior of biological molecules. Genes, the basic elements of biological information, reflect the molecular structures of the enormously large, linear DNA molecules of which they are made. The behavior of enzymes, hormones, and antibodies reflects the molecular structures of proteins and the organic chemistry of the functional groups of the amino acid side chains. The surface and barrier properties of biological membranes reflect the ability of lipids to aggregate into flexible two-dimensional bilayers with hydrophobic cores and polar surfaces.

Information about the molecular structures and biophysical properties of proteins, nucleic acids (DNA and RNA), lipids, and carbohydrates is available on the Molecular Structure

The characterization of molecular structure, the measurement of molecular properties, and the observation of molecular behavior presents an enormous challenge for biological scientists. A wide range of biophysical techniques have been developed to study molecules in crystals, in solution, in cells, and in organisms. These biophysical techniques provide information about the electronic structure, size, shape, dynamics, polarity, and modes of interaction of biological molecules. Some of the most exciting techniques provide images of cells, subcellular structures, and even individual molecules. It is now possible, for example, to directly observe the biological behavior and physical properties of single protein or DNA molecules within a living cell and determine how the behavior of the single molecule influences the biological function of the organism.

Information about the wide variety biophysical techniques available to study the structures, properties, and functions of molecules both in the test tube and in living biological systems is available on our Biophysical Techniques page.

Much of the scientific success of computational biology depends upon its ability to develop detailed physical mechanisms to explain specific biological processes. The double helical structure of DNA, for example, provides a framework for an explanation of how genetic material is replicated and of how genetic mutations arise: specific proteins mediate the unwinding of the DNA duplex and the assembly of a new strand based on complementary base pairing of the four DNA bases, guanine with cytosine and adenine with thymine; mismatch of one of these base pairs generates a complementary strand with a single base substitution. The value of this, and a variety of other biophysical mechanisms, is unlimited for human knowledge in general and for biomedical research in particular.

Development of high throughput data acquisition technologies in biological sciences, together with advances in digital storage, computing, and communications technologies have begun to transform biology in general, and molecular biology in particular, from a data poor science to a data rich science. Examples of biological data sets include DNA and protein sequence data, macromolecular structure and function data, and gene expression data.

A major goal of computational molecular biology is to develop information processing or algorithmic accounts of biological questions such as: How is information encoded, stored, decoded, and used in biological systems? What sequence regularities (if any) are predictive of protein function? How can we precisely characterize the syntax (grammar) and semantics (meaning) of macromolecular sequences? How do hundreds of genes interact over time to orchestrate specific biological processes of interest (neural development, disease, aging)? Research in computational biology requires the development of sophisticated databases, knowledge bases, ontologies, algorithms and software tools for data storage and retrieval, data integration, information extraction, exploratory data analysis and discovery (through data mining and data visualization), experiment design, using heterogeneous biological data sources. Design and development of such tools is a major goal of bioinformatics or genome informatics.

All in all two distinct branches can be distinguish in computational biology: knowledge discovery, or data-mining, which extracts the hidden patterns from huge quantities of experimental data, forming hypotheses as a result; and simulation-based analysis, which tests hypotheses with in silico experiments, providing predictions to be tested by in vitro and in vivo studies.

Knowledge discovery is used extensively within bioinformatics for such tasks as the prediction of exonintron and protein structure from sequence, and the inference of gene regulatory networks from expression profile. These methods typically use predictions based on heuristics, on statistical discriminators that often involve sophisticated approaches (such as hidden Markov models) and on other linguistic−based algorithms.

In contrast, simulation attempts to predict the dynamics of systems so that the validity of the underlying assumptions can be tested. Detailed behaviours of computer−executable models are first compared with experimental observation. Inconsistency at this stage means that the assumptions that represent our knowledge on the system under consideration are at best incomplete. Models that survive initial validation can then be used to make predictions to be tested by experiments, as well as to explore questions that are not amenable to experimental inquiry.

Although traditional bioinformatics has been used widely for genome analysis, simulation−based approaches have also received mainstream attention. Current experimental molecular biology is now producing the high-throughput quantitative data needed to support simulation−based research. Combined with rapid progress of genome and proteome projects, this is convincing increasing numbers of researchers of the importance of a system level approach. At the same time, substantial advances in software and computational power have enabled the creation and analysis of reasonably realistic yet intricate biological models.

Recent years have witnessed an exponential growth of data on the 3D structures of macromolecules, and in particular proteins. Managing this information is a challenging problem. It requires efficient ways of storing, cross referencing and accessing these data and the in formation that can be obtained from them, commonly re ferred to as *databases.* Such databases can only be useful if the data they contain are consistent and as error free as possible. This applies in particular to the atomic coordinates of the macromolecules and genes data sequencing.

It is clear that results of computational biology deeply rely on the quality of experimental data. This is particularly true for those methods that heavily make use of experimental data, such as large sequencing projects, protein-protein interactions network mapping or protein three-dimensional structure determination via both NMR or X-ray spectroscpy. Evaluating and assessing the quality of experimental data and refining experimental high-throughput methods should be a major goal of the paramount importance to be pursuived.

Chapter 1

Evaluation of Single Nucleotide Polymorphisms data quality

1.1 Introduction

If one compares the same stretch of DNA for two or more individuals, nucleotide polymorphisms are the most frequently observed differences at the nucleotide level within diploid organisms. Single nucleotide polymorphisms (SNPs) shown to be essential for studying a variety of organismal properties and processes, such as recombination, chromosomal dynamics, genome rearrangement, and genetic relatedness between individuals.

SNPs can be located in coding regions of genes or in intergenic regions, where they are most abundant [31], [70]. In coding regions, SNPs can alter the function and structure of encoded proteins, e.g., proteins involved in drug metabolism [77], [117]. In humans, single-nucleotide substitutions are the cause for most of the known recessively or dominantly inherited monogenic disorders, and missense SNPs also often contribute to common diseases [85] , [128]. SNPs are estimated to occur once every 1 kb throughout the human genome and are being targeted for association mapping of disease susceptibility genes [111] and used to study traits of diseases such as cancers, which are often accompanied by a loss of heterozygosity at SNP loci.

Recent surveys of human genetic diversity have estimated that there are about 250,000-400,000 common single nucleotide polymorphisms (SNPs) in protein coding sequences of the genome [29], [55]. Analysis of their functional effects is a crucial aspect of current genomic science. Coding SNPs (cSNPs) are interesting, in part, because some of them, termed non-synonymous SNPs (nsSNPs), introduce amino acid polymorphisms into their encoded proteins. nsSNPs are proportionally less prevalent than synonymous SNPs that do not affect protein sequence, presumably as a consequence of selection against the functional disruptions of amino acid variation [29], [55]. However, it might be expected that a significant fraction of molecular functional diversity in the human population remains attributable to effects on protein function caused by nsSNPs. For example, the kinetic parameters of enzymes, the DNA-binding properties of proteins that regulate transcription, the signal transduction activities of transmembrane receptors, and the architectural roles of structural proteins are all susceptible to perturbation by nsSNPs and their associated amino acid polymorphisms. Amino acid polymorphisms can also influence the efficacy and toxicity of drugs, as has been shown for cytochrome P450 2D6, TPMT, and the β_2-adrenergic receptor among others. [80], [79], [82], [132], [45], [116].

Structural analysis of amino acid polymorphisms provides a powerful mechanistic explanation of their effects on function. Very early in the molecular analysis of genetic variation, the strengths of structural analysis were demonstrated for the case of amino acid mutations in hemoglobin. Here, the molecular basis of the clinical effects caused by mutations could be inferred as soon as the structural information became available [103], [152].

These pioneering studies recognized crucial links between the structural disposition of residues and potential effects of mutations on function, including the destabilizing effects of introducing charged residues into the hydrophobic core of a protein, and the functional disruptions of mutations in protein residues that contact the iron or the heme ligands.

As the structure databases have grown and been analyzed by computational methods, the understanding of the relationship between structure and the effects of amino acid substitution on function has continued to deepen. Many studies have shown that a model residue's solvent accessibility is important for anticipating whether its mutation will affect function [22], [41]. In a similar way, a strong relationship between molecular rigidity measured by a crystallographic B-factor and the tolerance to mutation for the case of lysozyme has been demonstrated [5]. Several groups have noted a systematic intolerance to mutation in residues that are either extremely conserved in phylogeny or confined in their identity to particular classes of amino acid residues [89], [105], [129].

Others point to relationships between functional effects and hydrophobicity or residue volume [22]. Sunyaev et al. have started to examine the relationship between structural features and either human disease causing mutations or common human nsSNPs [131]. They identified structural features that are significantly associated with the disease causing polymorphisms, and found that a surprisingly large fraction, about 45%, of the prevalent nsSNPs share these structural features. For about half of these structurally impor-

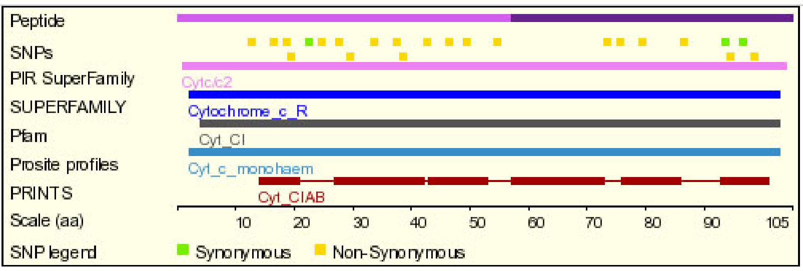


Figure 1.1: Graphical display of human cytochrome *c* gene features. Source: `ensemble.org`

tant residues, the polymorphism represents an amino acid substitution that apparently is not found in interspecies variation.

The current number of known protein structures is still far less than the number of known human protein sequences, but this discrepancy does not diminish the importance of structural analysis for understanding the effects of nsSNPs and their amino acid polymorphisms on function. It is accepted generally that proteins with similar amino acid sequence will exhibit a high degree of structural homology, even when they are only distantly related. For example, in hemoglobin and myoglobin (sharing only 25% amino acid identity in sequence) the structural dispositions of many corresponding residues are extremely well conserved. The same principle serves as the underlying basis of classifications of proteins according to fold families (e.g. DALI, SCOP) [56], [94] and structure prediction methods like threading [100], [1] and homology modeling [52], [118].

Shared structural properties can be extremely precise, for example in the conservation of the residues that coordinate the heme in hemoglobin and myoglobin. They can also be more general, as for example if the corresponding residues in two proteins are both hydrophobic and buried in the hydrophobic core. For the majority of proteins, structural information is not available; but for those proteins with sequence homology to a protein of known structure, much structural information can be inferred. For the human genome, about 30% of the protein sequences are likely to be homologous to known crystal or NMR structures [52]. The current high-throughput structure initiatives and theoretical modeling techniques will increase this proportion dramatically in the next years [2], [8], [4], [51].

In 1999, Buetow et al. [26] and in 2000 Irizarry et al. [61] published two papers in Nature Genetics proposing two computational strategies to identify candidate SNPs in expressed sequences from public EST data.

An expressed sequence tag or EST is a short sub-sequence of a transcribed

Residue	**dbSNP rs ID**	**SNP type**	**Handle|Submitter ID**
14	rs11548821	ns	CGAP-GAI
18	rs17851278	ns	YMGC ENOME DIFF
20	rs11548816	ns	CGAP-GAI
21	rs11548815	ns	CGAP-GAI
24	rs11548797	s	CGAP-GAI
26	rs11548802	ns	CGAP-GAI
29	rs11548799	ns	CGAP-GAI
31	rs11548796	ns	CGAP-GAI
35	rs11548805	ns	CGAP-GAI
39	rs11548791	ns	CGAP-GAI
40	rs11548812	ns	CGAP-GAI
44	rs1154880	ns	CGAP-GAI
48	rs11548783	ns	CGAP-GAI
51	rs11548772	ns	CGAP-GAI
56	rs11548795	ns	CGAP-GAI
75	rs11548820	ns	CGAP-GAI
77	rs11548818	ns	CGAP-GAI
81	rs11548778	ns	CGAP-GAI
88	rs11548785	ns	CGAP-GAI
95	rs3211448	s	LEE
96	rs3211449	ns	LEE
98	rs11548776	s	CGAP-GAI
100	rs3211451	ns	LEE

Table 1.1: List of human cytochrome *c* polymorphisms deposited in dbSNP build 127.

spliced nucleotide sequence (either protein-coding or not). They may be used to identify gene transcripts, and are instrumental in gene discovery and gene sequence determination [3]. The identification of ESTs has proceeded rapidly, with approximately 43 million ESTs now available in public databases as GeneBank database [14].

An EST is produced by one-shot sequencing of a cloned mRNA (i.e. sequencing several hundred base pairs from an end of a cDNA clone taken from a cDNA library). The resulting sequence is a relatively low quality fragment whose length is limited by current technology to approximately 500 to 800 nucleotides. Because these clones consist of DNA that is complementary to mRNA, the ESTs represent portions of expressed genes.

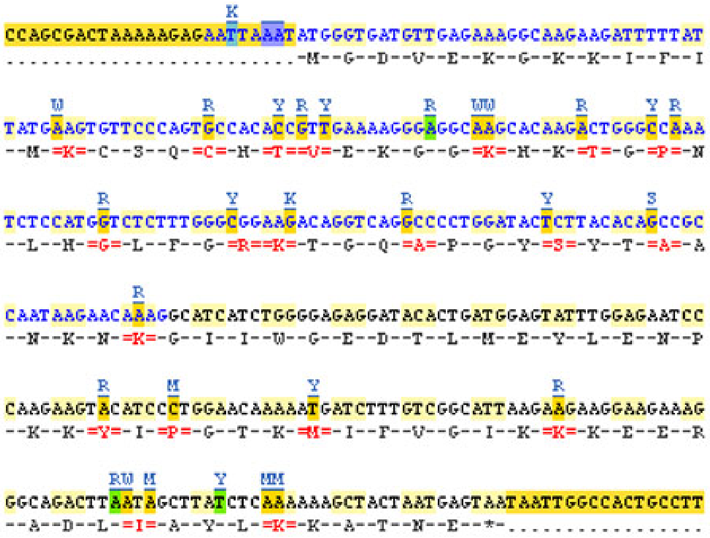
The current understanding of the human set of genes includes the existence of thousands of genes based solely on EST evidence. In this respect, ESTs become a tool to refine the predicted transcripts for those genes, which leads to prediction of their protein products, and eventually of their function. Moreover, the situation in which those ESTs are obtained (tissue, organ, disease state - e.g. cancer) gives information on the conditions in which the corresponding gene is acting [122]. Anyway EST are intended solely to find a consensus sequence for a gene [36] and not to detect differences from an assessed sequence. A detailed review of EST methods can be found in [97].

As a result of their study, Buetow et al. and Irizarry et al. added a significant number of SNPs to the newborn dbSNP database [120]. Other EST−based computational methods have since been developed and new SNPs identified by these methods are being deposited on a regular basis. As of July 2006 (dbSNP build 127), more than 16% of all SNPs in coding sequences (cSNPs) originate from these computational methods.

These EST-based dbSNP data are used with the same level of confidence as all other SNPs. Indeed, the original papers claimed an overall success rates of 70%÷80% through sample validation experiments. However, an internal check within the dbSNP itself shows that only 28.2% of the cSNPs deriving from EST-based computational methods have been detected also by experimental methods. This success rate is already unexpectedly low, and we find that it decreases further (24.7%) when the comparison is made more stringent by restricting it to the genes analysed in the frame of the NIEHS Enviromental Genome Project (EGP) [84] (`http:\\egp.gs.washington.edu`). In this comparison, the Buetow et al. cSNP dataset scores as low as 19%, while the Irizarry et al. cSNP dataset scores 33%.

To confirm or disprove these more pessimistic estimates, we decided to experimentally verify the presence of cSNPs in a single gene on a statistically meaningful sample of different individuals, and we selected the same pool of individuals used in the EGP project. To increase the significance of the test we looked for - relatively small - genes possessing a relatively large number of candidate cSNPs in the coding region (see Table 1.1 and Figure 1.2).

The search pointed to somatic human cytochrome c (cyc *c*), with an ORF of 318 bp and 24 candidate cSNPs (6 synonymous and 18 non-synonymous), 15 of which were deposited in 2003 using the method of Buetow et al. [26]. The coding parts of the two cyc *c* coding exons and the short intervening intron were PCR-amplified from the selected panel of 95 individuals and bi−directionally sequenced. In such a way, from the number of experimentally verified cSNPs we expected to be able to provide a reasonable estimate of the reliability of the candidate cSNPs for this gene. Surprisingly, none of the 19 cSNPs was found in the 95-individuals panel (190 chromosomes). No other SNPs were found. Results of this study have been published in [18]. As



```
                   K
CCAGCGACTAAAAAGAGAATTAAATATGGGTGATGTTGAGAAAGGCAAGAAGATTTTTAT
.........................-M--G--D--V--E--K--G--K--K--I--F--I

    W            R     Y R Y       R    WW       R     Y R
TATGAAGTGTTCCCAGTGCCACACCGTTGAAAAGGGAGGCAAGCACAAGACTGGGCCAAA
--M-=K=--C--S--Q-=C=-H-=T==V=-E--K--G--G-=K=-H--K-=T=-G-=P=-N

        R          Y    K         R           Y        S
TCTCCATGGTCTCTTTGGGCGGAAGACAGGTCAGGCCCCTGGATACTCTTACACAGCCGC
--L--H-=G=-L--F--G-=R==K=-T--G--Q-=A=-P--G--Y-=S=-Y--T-=A=-A

           R
CAATAAGAACAAAGGCATCATCTGGGGAGAGGATACACTGATGGAGTATTTGGAGAATCC
--N--K--N-=K=-G--I--I--W--G--E--D--T--L--M--E--Y--L--E--N--P

        R     M           Y                    R
CAAGAAGTACATCCCTGGAACAAAAATGATCTTTGTCGGCATTAAGAAGAAGGAAGAAAG
--K--K-=Y=-I-=P=-G--T--K-=M=-I--F--V--G--I--K-=K=-K--E--E--R

         RW M     Y   MM
GGCAGACTTAATAGCTTATCTCAAAAAAGCTACTAATGAGTAATAATTGGCCACTGCCTT
--A--D--L-=I=-A--Y--L-=K=-K--A--T--N--E--*-.................
```

Figure 1.2: Human cytochrome *c* coding sequence together with the transcript sequence. Single nucleotide polymorphisms are highlighted. Source: `ensembl.org`. Ensembl access code for cyc *c* gene is ENSG00000172115

many methods and findings used in the study have not been fully reported in the paper, they will be extensively explained and discussed hereafter.

1.2 Material and Methods

1.2.1 Statistical Methods

Some general definitions

Given a set of genes with genotyped cSNPs, we indicated with s_i the number of SNPs in the coding region of length λ_i of the i-th gene. We get the average snp number η as

$$\eta = \frac{\sum_i s_i}{G} \tag{1.1}$$

where G is the total number of considered genes. We define SNP density ρ as

$$\rho = \frac{s_i}{\lambda_i} \tag{1.2}$$

We get the average snp density $<\rho>$ as

$$<\rho>= \frac{\sum_i \frac{s_i}{\lambda_i}}{G} \tag{1.3}$$

A statistical model for SNPs distribution

Let have a number S of putative Single Nucleotide Polymorphisms (pSNP). Let be f_i the frequency of the i-th pSNP. We can set the probability p_i to find the i-th pSNP in an individual as $p_i = f_i$. The probability $A(\mathbf{p}; S)$ to find zero pSNP can be estimated as:

$$A(\mathbf{p}; S) = \prod_{i=1}^{i=S} (1 - p_i) \tag{1.4}$$

where $\mathbf{p}$ is the vector of the $p_1 \ldots p_S$ pSNPs probabilities. Let us now consider a population of N individuals and a sample of dimension K. The probability $B(\mathbf{p}; S, K, N)$ to get K individuals with none of the S pSNPs is given by the equation

$$B(\mathbf{p}; S, K, N) = \binom{N}{K} \left(\prod_{i=0}^{i=S} (1 - p_i) \right)^K \left(1 - \prod_{i=1}^{i=S} (1 - p_i) \right)^{N-K} \tag{1.5}$$

Statistical tests for non normality of data

Deviation of an experimental distribution (*i.e.* coding and non coding SNPs densities) from normality can be assessed by mean of fundamental parameter such as *skewness* and *kurtosis* that can been tested using the statistical tests proposed by Snedecor and Cochran [123], . The symmetry of the distribution is given by the skewness m_3, defined as the third moment about the mean

$$m_3 = \frac{\sum_{i=1}^{n} (x_i - \mu)^3}{n} \tag{1.6}$$

The skewness is 0 if the distribution is symmetric, positive (negative) if the it is asymmetric on the right (left). Nevertheless, the absolute value of this parameter is strongly dependent on the scale utilized for the measurement. Said σ the standard deviation, adimensionalization is achieved by dividing it by σ^3. Following this approach, one get the Pearson β_1 index and the Fisher γ_1 index[1] defined as

$$\beta_1 = \left(\frac{m_3}{\sigma^3}\right)^2 \tag{1.7}$$

$$\gamma_1 = \frac{m_3}{\sigma^3} \tag{1.8}$$

with the relation

$$\gamma_1 = \sqrt{\beta_1} \tag{1.9}$$

The kurtosis m_4 (or *peakedness* or *tailed-ness*) is defined as

$$m_4 = \frac{\sum_{i=1}^{n}(x_i - \mu)^4}{n} \tag{1.10}$$

As kurtosis for a gaussian distribution is 3, Pearson β_2 and Fisher γ_2 indexes are defined as

$$\beta_2 = \frac{m_2}{\sigma^4} \tag{1.11}$$

$$\gamma_2 = \frac{m_2}{\sigma^4} - 3 \tag{1.12}$$

Significativity for both skewness and kurtosis can be assessed using a bilateral test if γ_1 or γ_2 indexes are not zero or a unilateral test if there is symmetry to the right (left) of if the distribution is platicurtic (leptocurtic), introducing a Z values [2] defined as (for big samples)

$$Z_{skewness} = \frac{\gamma_1}{\sqrt{\frac{6}{n}}} \tag{1.13}$$

$$Z_{kurtosis} = \frac{\gamma_2}{\sqrt{\frac{24}{n}}} \tag{1.14}$$

[1] Statisticians use to indicate Pearson and Fisher indexes as **b** and **g** when these parameters are calculated on an experimental distribution, but we will maintain the greek notations.

[2] These relations are derived from the general formula $Z = \frac{t-\mu}{\sigma}$. For big samples, one finds that γ_1 and γ_2 are distributed in a quasi-gaussian manner with standard deviation σ equal to $\sqrt{\frac{6}{n}}$ and $\sqrt{\frac{24}{n}}$ respectively.

Chi-square test for 2 × M tables

To asses difference among groups we use a χ^2 (Chi-square) test for $2 \times M$ tables using the exact formula:

$$\chi^2 = \sum_{i=1}^{2\times M} \frac{(f_i^{obs} - f_i^{exp})^2}{f_i^{exp}} \tag{1.15}$$

where f_i^{obs} are the observed frequencies and f_i^{exp} the expected ones

Tests for outliers

The more general method to estimate the probability that a particular data belongs to certain population, without *a-priori* knowledge or hypothesis on the shape of the data distribution is the so called Chebyshev inequality. Chebyshev inequality states that for a set of data the percentage P of observations comprised in the distance of k standard deviations σ from the mean μ is at least

$$P\left(k\frac{|t-\mu|}{\sigma}\right) \leq \left(1 - \frac{1}{k^2}\right) \tag{1.16}$$

independently from the distribution itself.

1.2.2 Bioinformatic methods

Databases surveying

Surveying the dbSNP database:

As for today, dbSNP127 counts 11.883.685 polymorphisms for *Homo sapiens* genome. Of these 6.262.709 (52%) have been validated.

The public SNPs database dbSNP1 (build 125) was queried using either the ENTREZ SNP filter mask or queries with Boolean operators. The success rate of candidate cSNPs derived from in silico analysis of public EST was estimated by restricting the queries to the panel of 546 genes (listed below) re-genotyped by the NIEHS Environmental Genome Project (EGP) [84]. EGP-derived SNPs and SNPs coming from the two EST-based computational methods published in Nature Genetics were retrieved in dbSNP according to the criterion that they had the [$HANDLE$] tags `EGP SNPS` for the NIEHS project dataset, `CGAP-GAI` for the Buetow et al.3 dataset (query restricted to `computed` [$METHOD$]) and `LEE` for the Irizarry et al.4 dataset. Experimental cSNPs were retrieved in dbSNP according to the criterion that they had the following [$METHOD$] tags: `sequence`, `hybridize`, `rflp`, `sscp`, `dhplc`.

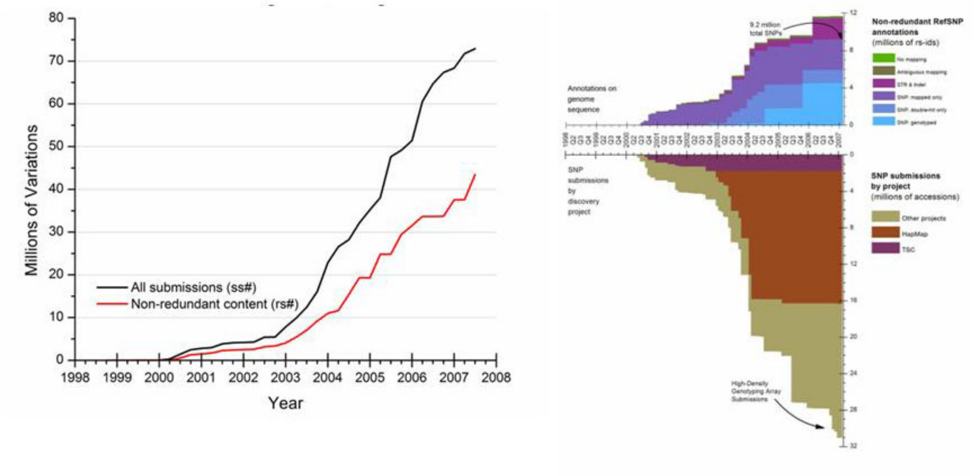


Figure 1.3: Growth of dbSNP for all organisms through end of August 2007 (left). The black line indicates the growth of dbSNP using the total number of submissions, while the red line indicates the growth using the non-redundant content (refSNPs) of dbSNP. Overall growth of human variations in dbSNP over the years since its inception (right). Through February, 2007 dbSNP is closing on 12 million non redundant variations clustered over 30 million submissions.
Source: `ncbi.nlm.nih.gov/SNP/`

Computational cSNPs were retrieved in dbSNP according to the criterion that they had `computed` $[METHOD]$ tag. All queries were restricted to cSNPs and to `snp` $[SNP\_CLASS]$.

SNPs number for the genes resequenced in the frame of the NIEHS project together with the approved HGCN names have been retrieved parsing the downloadable data sets available at the address `http:\\egp.gs.washington.edu\`. Genes coding regions length have been retrieved automatically parsing on-line the NCBI Entrez Gene database [87], [88]. An html page parser have been implemented to retrieve useful information from the pages `www.ncbi.nlm.nih.gov\sites\entrez?db=gene&cmd=search&term=HGNC` (where HGNC stands for the HGNC gene name [87]). In particular the associated expressed protein codes NP have been retrieved and linked to the deposited FASTA format sequences [101] through the link `www.ncbi.nlm.nih.gov\entrez\viewer.fcgi&val=NP&dopt=fasta` (where NP stands for the NP protein code). FASTA sequences have been used do retrieve univocally the length of the codding regions.

Software

The entire nucleotidic coding region of human cyc *c* aligned against the BLAST EST database [21] using the on-line version of the BLAST program [6] (`0-www.ncbi.nlm.nih.gov.catalog.llu.edu\BLAST\`). Default setting have been used.

Sequence alignment of re-sequenced cyc *c* has been performed with th on-line version of the program CULSTALW [133] (`www.ebi.ac.uk\Tools\clustalw\`).

In-house utilities for data mining, parsing and analysis have been implemented using Python programming language (`www.python.org`) and the MATLAB software from MathwWorks (`www.mathworks.com`).

Calculations have be performed under Linux OS (SuSE distribution 9.1, `www.opensuse.org`)

The freely available program `finchtv` was used for manual inspection of sequence cromathograms (`www.geospiza.com\finchtv\`).

1.2.3 Molecular Methods

PCR and sequencing

The CYC was amplified from genomic DNA of 95 samples from Coriell panel using forward (FW) primer 5'−AGTGGCTAGAGTGGTCATTCATTT ACA−3' and reverse (RW) primer 5'−TCATGATCTGAATTCTGGTGTAT GAGA− 3'. Amplification was performed in a 25 μL reaction volume using PfuTurbo DNA Polymerase (Stratagene). The amplified DNA fragments were analyzed by electrophoresis on a 1.5% agarose gel and purified with QI-Aquik PCR Purification Kit (Qiagen). Subsequently the amplified and purified DNA fragments were sequenced using both foward and reverse primers.

1.3 Results and Discussion

The first remarkable fact is the surprisingly high number of non-synonymous SNPs in a so quite short coding region. Human cyc *c* gene transcript is 5506 bps long. Coding regions is 318 bps, 5'Utr is 146 bps and 3'Utr is 5042 bps. Translated protein is 106 residues long (counting start Methionine).

Using relation (1.2) and dbSNP126 data we get for human cyc *c* a nsSNP density $\rho_{nsyn} = 0.05$ snp/bps and a sSNP density $\rho_{syn} = 0.009$ snp/bps.

We did compare these values with average values of a large set of human genes with experimentally validated SNPs. From NIEHS SNPs Program database we found 388 genes with genotyped and experimentally validated SNPs: human cyc c was not been sequenced. We found that the average number η_{nsyn} of nsSNPs in coding region is $\eta = 3.7 \pm 3.7$ while the number η_{syn} of sSNPs in coding region is $\eta = 3.7 \pm 3.4$.

Using equation (1.3) we get $<\rho_{nsyn}>= 0.006 \pm 0.005\,\text{snp/bps}$ for nsSNPs and $<\rho_{syn}>= 0.006 \pm 0.006\,\text{snp/bps}$. These results agreed with values found by Zhao et al. [149], [150].

From these data appears that human cyc c has a number of nsSNPs greater than η_{nsyn} of over 3 standard deviations and a nsSNP density greater than 7 standard deviations in respect to $< \rho_{nsyn} >$. The number of sSNPs and sSNPs density value both lay in error ranges.

The distributions for ρ_i values are highly asymmetric and not gaussian shaped: $m_3 = 1.2$, $m_4 = 3.9$ for nsSNP, $m_3 = 0.8$, $m_4 = 4.4$ for sSNP. The Z values for Fisher γ_1 and γ_2 indexes (equations (1.8) and (1.11)) for skewness and kurtosis tests are both $> 10^6$ and non-normality of the two distribution is highly statistically significant ($\alpha << 10^{-4}$).

To estimate if nsSNP density value for cyc c is an outlier from the distribution of nsSNPs densities, we used relation (1.16) as data are not normally distribuited. The probability that $\rho_{nsyn} = 0.05\,\text{snp/bps}$ is an outlier from distribution is greater than 98.6%.

These statistical considerations paired with the structural and functional impact of some of the reported mutations (for a discussion see [18]) drive us to search experimental validation (i. e. Minor Allele Frequency) for reported putative SNPs. Sequencing of 95 Dna samples and alignment of all sequences did not show the presence of SNPs in homozygotes, while manual analysis of chromatograms did not allow to find SNPs in heterozygotes. As a confirmation, we also draw some considerations about probability.

Allele frequencies or heterozygosity for SNPs deriving from computational methods are obviously not available as these parameters make sense only if derived from an experimentally genotyped population. Confounding probability with frequency we did operatively assumed $p_i = f_i$; supposing to have to deal with very rare SNPs (as the lower accepted threshold value of the sequence for common variants is $f = 0.01$) we set $p_i = p = 0.01$ for all the putative SNPs. With this assumption, equations (1.4) and (1.5) become

$$A(p;S) = (1-p)^S \tag{1.17}$$

$$B(p;K,N) = \binom{N}{K} (1-p)^{SK} \left[1-(1-p)^S\right] \tag{1.18}$$

We did sequence $K = 95$ individuals on a panel of $N = 95$, looking for $S = 24$ cSNPs. With these values, equation (1.4) gives a probability to get an individual with no pSNPs $A = 0.79$ (79%); equation (1.5) reduces to $B = A^N$ giving $B = 1.2 \times 10^{-10}$. That is is really unlikely to find 95 without SNPs by chance. This show that also if dealing with really rare SNPs the probability of finding none of them in 95 individuals is really low.

To better understand the meaning of this negative result we tried to reproduce the results of [26] and [61], blasting entire nucleotidic coding region of human cyc c against the BLAST EST database. On over 1000 matching sequences, 619 sequences have 100% identity with the query sequence while the remaining have sequence indentity lower than 100%. This was due substiantally to base mismatching sequences (104) or to gapped sequences (201). From the alignement of the mismatching sequences and the cyc c reference sequence we found tha 100 possible variations in EST sequences resulting in 93 putative non synonymous SNPs and 26 synonynmous SNPs. We did notice that many of mutated sequences come from tumoral tissues as well for the matching sequences. Of 104 mismatching sequences, 48% originated from a tumoral tissue, 34.6% from normal different tissues and 17.4% form unknown tissues. Of the 619 matching sequences 50% originated from a tumoral tissue, 34.6% from normal different tissues and 15.4% form unknown tissues. We then asked ourselves if the high number of SNPs in the cyc c could arise from the fact that also many EST used also by Buetow et al. were from non normal tissues. We did try to associate SNPs presence to tumoral tyssue but A χ^2 test ($M = 3$, $d.o.f. = 2$, $\chi^2 = 9.067$, $\alpha = 0.01$, $\chi^2_{crit} = 9.210$ and $\alpha = 0.005$, $\chi^2_{crit} = 10.597$) shows that the difference beetwen the three groups is not statistically significant. That is no inference can be assessed beetween SNPs frequency and tissue type.

Several of the nonsynonymous mutations involve residues that have been shown to play crucial roles in the stabilization of the protein core and its solvation and in maintaining the conformational integrity of the molecule and of the surface involved in protein protein interactions. The reported mutations rs11548778 (Met81Thr) and rs11548796 (Pro31Ser) are both in absolutely conserved residues: Met81 is involved in the coordination of the iron ion of the heme prosthetic group, while Pro31 is involved in hydrogen bonding of the imidazole ring of the coordinated histidine. The iron ion of the heme is six-coordinated: two of the binding atoms are the Hδ1 atom of the side chain of His19 and the sulfur atom of the side chain of Met81. The Hϵ1 proton of the imidazole is hydrogen bonded to the carbonyl oxygen of Pro31. The nsSNP rs17851278 (Cys18Tyr) again occurs in a totally conserved position surrounding the so-called heme pocket, a region that is known to have both structural and functional roles. The nsSNP rs11548785 (Lys88Arg) affects

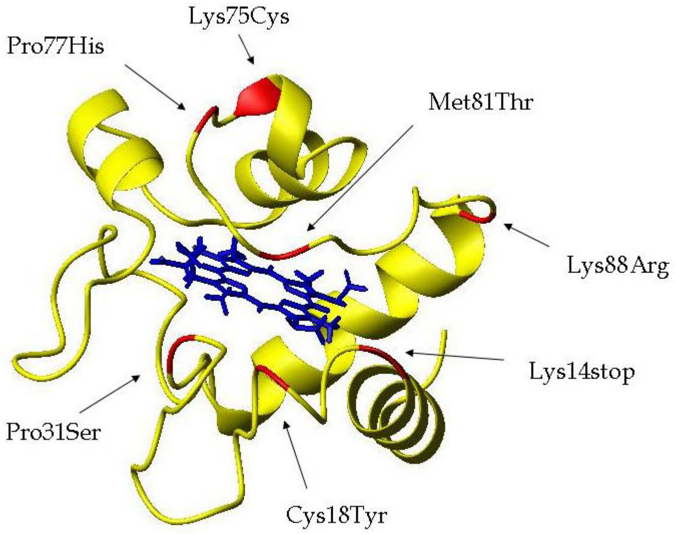


Figure 1.4: NMR structure of the reduced recombinant human cytochrome *c* (PDB entry 1J3S [64]). Residues affected by critical putative nsSNPs are highlighted in red.

one of the contact sites involved in the cyc *c*-ATP interaction. The stop-gain mutation rs11548821 (Lys14stop) occurs in a conservatively substituted position (only Arg or Lys are found in all known species) that has been proposed to be involved in the interactions of the complex between cytochrome *c* and the cytochrome *c* peroxidase [102]. Other nonsynonymous mutations, such as rs11548820 (Lys75Cys) and rs11548818 (Pro77His), occur in a very highly conserved region (>95%). In total, 39% of cyc *c* putative nonsynonymous SNPs are in highly or totally conserved regions with predicted or demonstrated functional significance [10].

1.4 Conclusions

Single nucleotide polymorphisms (SNPs) are important tools in studying complex genetic traits and genome evolution. Computational strategies for SNP discovery make use of the large number of sequences present in public databases like expressed sequence tags. These methods are considered to be faster and more cost-effective than experimental procedures. A major challenge in computational SNP discovery is distinguishing allelic variation from sequence variation between paralogous sequences, in addition to recognizing sequencing errors. For the majority of the public EST sequences, trace or quality files are lacking which makes detection of reliable SNPs even more difficult because it has to rely on sequence comparisons only.

From our re-examination of the SNPs deposited in dbSNP by EST-based methods we can conclude that on average only one fourth of the candidate cSNPs are true SNPs and this casts serious doubts on the validity of this kind of computational approach.

Chapter 2

Refining Yeast Two-Hybrid methods

2.1 Introduction

Protein-protein interactions have attracted much attention because they form the basis of a wide variety of biochemical reactions. The identification of proteins that interact with a known protein is an essential aspect of the elucidation of the regulation and function of that protein.

Many proteins mediate their biological function through protein interactions, so the systematic identification of such interactions for a given proteome has been proposed as a potentially informative functional genomic strategy [13], [73], [145].

As well as their well-described role in the assembly of a cell's structural compartments such as the cytoskeleton and nuclear pore, protein interactions are crucial for many other aspects of cellular biology. First, genetic interactions often correlate with physical interactions between the corresponding gene products. For example, it has been shown that in a *C. elegans* pathway that regulates apoptosis, the *ced-3* caspase gene acts downstream of *ced-4*, which itself acts downstream of *ced-9* [147]. Subsequently, the corresponding proteins CED-3, CED-4 and CED-9 were shown to interact physically [34]. Second, protein interactions are required to tether the components of signal-transduction pathways physically. A protein providing such a scaffold is the yeast protein Ste5p, which interacts with components of a mitogen-activated protein kinase (MAPK) cascade involved in pheromone signalling [35]. Third, enzyme protein substrate interactions are important for catalysis, and are often found to be more stable than previously presumed. For example, a cyclin-dependent kinase physically interacts with its substrate p107, a homologue of the retinoblastoma protein pRb [151]. Last, protein interactions are crucial for the integrity of multicomponent enzymatic machines such as RNA polymerases or the splicesome.

This interest has stimulated the development of a number of biochemical and genetic approaches to identify and clone genes encoding interacting proteins includind coimmunoprecipitation, copurification, cross-linking, and direct expression library screening using proteins as probe. However, the development of the yeast two-hybrid appears to have had the greatest impact on interaction cloning methodology.

2.2 Fundamentals of Yeast Two-Hybrid

The observation that most transcription factors can separated into a DNA-binding domanin (DB) and a transcriptional actovation domanin (AD) led to the development of the yeat tewo-hybrid (Y2H) system [49]: this system was devised to identify genes encoding proteins that phisically associate with a given protein in vivo. This is a versatile and powerfull method that is applicable to most, if not all, proteins once their genes have been isolated.

Trascriptional factors (TRs) that activate transcription are usually composed of a DNA-binding domain and an activation domain. The binding domain targets these proteins to a specific binding site in the promoter or enhancer region of a gene while the activation domanain mediates transcription initiation. Transcriptional activators in yeast were among the first to be studied in detail. In all known cases they recruit additional proteins or whole complexes to the pertinent promoters, eventually leading to the binding of one of the three RNA polymerases [137], [134].

While DBDs are extremely well characterized both functionally and structurally, ADs do not share easily recognizable motifs or structures. Accordingly, no specific pattern or motif for the identification of an AD has been defined in pattern/domain databases such as Prosite [60] or SMART [119], [76]. In contrast, more than 50 patterns for DBDs have been documented in the SMART database. While DBs are extremely well characterized both functionally and structurally, ADs do not share easily recognizable motifs or structures. Accordingly, no specific pattern or motif for the identification of an AD has been defined in pattern/domain databases. In contrast, more than 50 patterns for DBs have been documented.

Uisng the Y2H, one can identify potentially interacting proteins (X-Y heterodimers or X-X homodimers) by generating two different hybrid proteins: one with protein X fused to DB and the other with protein Y fused to AD. If the protein X and Y interact, the AD can be brought close to the

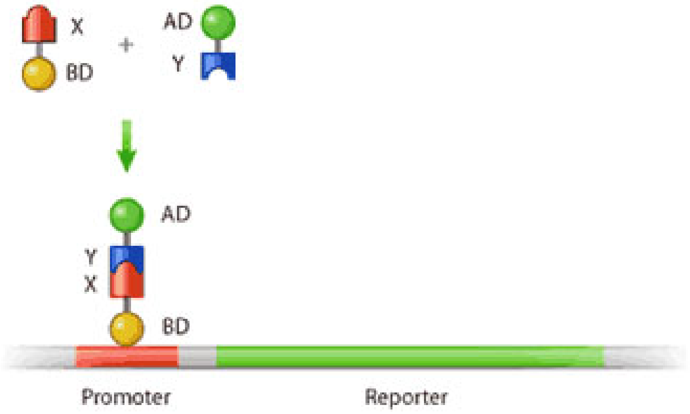


Figure 2.1: Two-hybrid interactions: two fusion proteins are created, DB-X and AD-Y each carrying a part of the transcriptor factor. If the protein X fails to interacct with Y, the the activation domain is not brought in proximity of the promoter and the reporter gene will be not activate. When proteins X and Y interact sucessfully, the activation domain is brought to the promoter that activate the reporter gene sucessfully.

promoter by DB-X and thereby activate the gene driven by that promoter that usually is a selectable or screenable marker (see Figure 2.1)

In contrast to biochemical methods detecting protein-protein interaction, this system is based on yeast genetic assay in which the interaction of two proteins is detected by the reconstitution of a functional transcription activator in yeast. This method not only allows identification of proteins that interact, but also can be used to define and/or test the domains/residues necessary for the interaction of two or more proteins [27], [44], [78], [28], [81], [114]. A test of the responsiveness of two different yeast strains to be used in (high-throughput) Y2H experiments was carried in the frame of a stage in Dr. Marc Vidal at Harvard Medical School. .

2.2.1 Reporter genes in yeast

Among the many different reporter genes integrated into the yeast genome are the *GAL1::HIS3, GAL1::lacZ, SPAL10::URA3, GAL2::ADA2* They contain similar, albeit non identical, Gal4p binding sites in their promoters. However, the sequences neighboring the Gal4p binding sites are different between the four reporters. The presence of different TATA boxes and initiation site sequences allow the discrimination between promoter context-specific reporter activation and two-hybrid interaction. Each reporter gene can be scored using a different two-hybrid phenotypes [142].

The *GAL1::HIS3* reporter gene is fused to its own promoter lacking endogenous Upstream Activation Sequence (UAS) sequence and containing 125 bp $GAL1_{UAS}$. *HIS3* encodes for the enzymes imidazole-glycerol-phosphate dehydratase required for the biosynthesis of histidine. The 3-amino-1,2,4-triazole (3-AT) can act as an inhibitor of this enxymatic activity [142] [40]. The expression of this gene is assayed by growth on media lacking histidine and containing 3-AT.

The *GAL1::lacZ* reporter consists in the *GAL1* full-length promoter fused to the *lacZ* gene which encodes for the β-galactosidase enzyme (β-gal). Activation of this gene can be assayed by colorimetric detection of the the β-gal using 5-bromo-4-chloro-3-indolyl- β-D-galactopyranoside (X-Gal) or orthonitrophenyl-β-D-galactopyranoside (ONPG). β-gal converts X-gal and ONPG to a blue or yellow product respectively. The intensity of the color reflects *lacZ* espression [93].

The *GAL2::ADA2* provides growth on media lacking adenine, as well as a range of color from red to white depending on the strength of Gal4p expression.

The *SPAL10::URA3* reporter gene is composed by the *SPO13* promoter with 10 Gal4p binding sites fused to *URA3*. In fact, most part of the Gal4p-inducible promoters in yeast show basal espression strong enough to confer $\text{Ura}^{+}\text{FOA}^{S}$ phenotype in absence of Gal4p. Thus a promoter was constructed that contains a *cis*-acting Upstream Repressing Sequence to mantain low levels of basal espression in cells lacking Gal4p, and a large number (10) of Gal4p binding sites to allow Gal4p-dependent activation. *SPAL10::URA3* is detected by growth on media lacking uracil or by the absence of growth on media containing both uracile and 5-fluoro-orotic acid (5-FOA). This enzyme, in fact, can convert the non toxic 5-FOA into the toxic compound 5 Fluorouridine monophosphate [143]. This is essential for the reverse two-hybrid system, in which mutations are selected that prevent an interaction [140], [141]. The reverse two-hybrid negative selections are based on the use of a reporter gene whose expression causes toxicity under specific growth conditions (counterselectable marker).

2.2.2 The MaV103/203 and Y8800/8930 yeast strains

S. cerevisiae strains MaV103 (*MAT**a***) and MaV203 (*MAT*α) contain deletions of the *GAL4* and *GAL80* genes encoding Gal4p and its repressor Gal80p, respectively. In the absence of Gal80p, galactose is not required for activation of Gal4p-inducible promoters. The strains contain three Gal4p in-

Auxotrophic markers	**Description**
leu2-3,112	Need for Leu in medium
trp1-901	Need for Trp in medium
his3Δ200	Need for His in medium
ura3-52	Need for Ura in medium
ade2-101	Reverted back to wild-type through an unknown mechanism
Other	**Description**
gal4Δ	Deletion of endogenous Gal4p transcription factor
gal80Δ	Deletion of Gal4p repressor.
$cyh2^R$	Resistance to Cycloheximide
$can1^R$	Resistance to Canavinine
Reporters	**Description**
GAL1::HIS3@LYS2	125 bp $GAL1_{UAS}$ in the *HIS3* promoter driving *HIS3*, integrated at the *LYS2* locus.
GAL1::lacZ	The *GAL1* promoter driving LacZ, integrated at an unknown locus.
SPAL10::URA3@ura3	*SPO13* promoter with 10 Gal4p binding sites driving *URA3*, integrated at *ura3-52* locus.

Table 2.1: Genotype of the MaV103/203 yeast strain for two-hybrid assay.

ducible reporter genes, providing four phenotypes to identify true interactors. The reporter genes are *HIS3*, providing for growth on plates lacking histidine, LacZ, for colorimetric detection of Gal4p activity, and *URA3*, providing growth on media lacking uracil, as well as sensitivity to the drug 5-FOA. In addition, the strains are resistant to canavanine and cycloheximide, aiding in plasmid shuffling. Both strain MaV103 and MaV203 show basal a espression of *HIS3* which is sufficient to allow growth on plates lacking histidine. Thus 3-AT can be titrated to provide a selection for moderate as well strong increases in *HIS3* espression [142]. MaV103 and MaV203 were derived from a cross between two non-isogenic strains PCY2 and MaV99 [33], [140].

S. cerevisiae strains Y8800 ($MAT\boldsymbol{a}$) and Y8930 ($MAT\alpha$) contain deletions of the *GAL4* and *GAL80* genes encoding Gal4p and its repressor Gal80p, respectively. In the absence of Gal80p, galactose is not required for activation of Gal4p-inducible promoters. The strains contain three Gal4p inducible reporter genes, providing four phenotypes to identify true interactors. The reporter genes are *HIS3*, providing for growth on plates lacking histidine, LacZ, for colorimetric detection of Gal4p activity, and *ADA2*, providing growth on

Auxotrophic markers	**Description**
leu2-3,112	Need for Leu in medium
trp1-901	Need for Trp in medium
*his3*Δ200	Need for His in medium
ura3-52	Need for Ura in medium
Other	**Description**
*gal4*Δ	Deletion of endogenous Gal4p transcription factor
*gal80*Δ	Deletion of Gal4p repressor
cyh2^R	Resistance to Cycloheximide
Reporters	**Description**
GAL2::ADE2	*ade2-101* locus replaced by *GAL2* promoter driven *ADE2*
GAL1::HIS3@LYS2	*GAL1* promoter driven *HIS3* gene downstream of *LYS2* gene
SPAL10::URA3@ura3	*SPO13 GAL7* promoter driven *LacZ* gene inserted in the *MET2* locus.

Table 2.2: Genotype of the Y8800/8930 yeast strain for two-hybrid assay.

media lacking adenine, as well as a range of color from red to white depending on the strength of Gal4p expression. In addition, the strains are resistant to cycloheximide, aiding in plasmid shuffling. Y8800 and Y8930 were generated by adding cycloheximide resistance to the PJ69-4 Y2H strains [62].

2.3 Material and methods

2.3.1 Materials

Yeast media

Synthetic Complete (SC) dropout medium: For MaV103/MaV203, use -Leu -Trp -His -Ura amino acid mixture. For Y8800/Y8930, use -Leu -Trp -His -Ade amino acid mixture. A -Leu -Trp -His -Ura -Ade is is suitable for both. For 11 L medium: 1.3 g aminoacids (AA) powder, 1.7 g Yeast Nitrogen Base (YNB,without AA and Ammonium Sulfate), 5 g Ammonium Sulfate. Adjust pH to 5.9. Autoclave to sterilize. Add 50 ml 40% glucose. Add 8 mL of each additional AA. (Leu, Trp, His, Ura, Ade) and 3AT or 5-FOA.

SC additives: 100 mM Histidine (filter to sterilize), 100 mM Leucine, 40 mM Tryptophan, 40 mM Adenine, 20 mM Uracil (filter to sterilize) 3AT concentrations range from 20-100 mM for MaV103/MaV203, and 0-5 mM for

Y8800/Y8930. 20 mM 3AT 1.68 g/l, 5-FOA 1 g/L or 2 g/L. Cycloheximide 1 μg/ml in ethanol.

YEPD: For 1 L medium: 20 g yeast peptone 10g yeast extract. Autoclave to sterilize. Add 50 ml 40% glucose. For *ade2* strains Y8800 and Y8930, the same amounts as for SC medium of extra adenine needs to be added.

Plate pouring: add 20 g od agar to 500 ml ddH$_2$O. Add 425 ml of the corresponding liquid media without supplements. Autoclave to sterilize. Add 50 ml 40% glucose and the desired supplements.

Solutions

1 M lithium acetate stock solution: Add 51 g of lithium acetate into 500 ml of ddH2O. Autoclave to sterilize.

10× TE: 100 mM Tris-HCl (pH 7.5) 10 mM EDTA, autoclave to sterilize.

50% PEG stock solution: Dissolve 125 g of polyethylene glycol (MW 350) in warm ddH$_2$O and finalize to 250 ml. Sterilize by filtration.

TE/LiAc: for 50 mll, add 5 ml of 10× TE and 5 ml of 1 M LiAc into 40 ml of sterile ddH$_2$O.

TE/LiAc/PEG: To make 50 ml, add 5 ml of 10× TE and 5 ml of 1 M LiAc into 40 ml of 50% PEG.

Boiled ssDNA: Boil the 10 mg/ml of salmon testes DNA for 5 to 10 min and chill on ice before transformation.

Z buffer: for 1 L 16.1 g $Na_2HPO_4 \cdot 7H_2O$, 5.5 g $NaH_2PO_4 \cdot H_2O$, 0.75 g KCl, and 0.246 g $MgSO_4 \cdot 7H_2O$. Autoclave to sterilize.

4% X-Gal: Dissolve 40 mg in 1 ml of N,N-dimethylformamide. Store at -20° wrapped with foil.

β-Gal solution: For each plate, 5 ml of Z buffer with 120 *mu*l of 4% X-Gal and 13 μl of 2- mercaptoethanol.

Lysis solution: for 1 ml, add 2.5 mg zymolyase to 0.1 M $MaPO_4$ buffer (pH7.4).

2.3.2 Yeast methods

Preparation of DB-ORF baits strains

The following protocol descirbes how to introduce DB (AD) ORF baits in yeast [144]. This protocol assumes using yeast strain MaV203. In case of Y8800/Y8930 strain adenine must be supplemnted to media (see 2.3.1) 1) Start an overnight culture of MaV203 yeast by scratching a small clump of cells from a patch into at least 0.5 ml of media for each bait you plan to transform (a minimum of 5 ml of media should be used). 2) The next day, take 0.5 ml of the overnight culture for each transformation. 3) Spin down the cells at 2000 rpm for 5 min. 4) Wash the cells by adding 0.25 ml of ddH$_2$O for each transformation. 5) Spin down the cells and wash in 100 ml of TE/LiAc for each transformation. 6) Spin down the cells and resuspend the cell pellet in 20 ml of TE/LiAc for each transformation. 7) Add 2 ml of boiled ssDNA for each tansformation. 8) Aliquot 22 ml of yeast into the wells a 96-well plate. 9) Add 50-100 ng of the appropriate DB-ORF DNA to each well. A transformation without DB-ORF DNA serves as a negative control. 10) Add 100 ml of TE/LiAc/PEG to each well and mix by pippetting. 11) Incubate at 30°C for 30 min. 12) Heat shock at 42°C for 15 min. 13) Spin down and remove the TE/LiAc/PEG solution with a multichannel pipette. 14) Add 120 ml of ddH$_2$O to each well, but be careful not to resuspend the cells. Remove 105 ml of ddH$_2$O from each well and resuspend the cells in the remaining 15 ml of liquid. 15) Spot 6-7 μl onto SC-Leu plates. 16) Incubate for 2 to 3 days at 30°C.

Introduction of an AD-cDNA library

The following protocol descirbes how to introduce a cDNA library in yeast cells [144], [78], [142]. This protocol assumes using yeast strain MaV203. In case of Y8800/Y8930 strain adenine must be supplemnted to media (see 2.3.1) 1) Grow the DB-ORF baits in 3 ml of SC-Leu yeast medium at 30° for approximately 24 ml. 2) Resuspend the cells well by vortexing and inoculate 10-15 ml in 35 ml of YPD and incubate for 15 to 18 h at 30° until the OD$_{600}$ reach 0.3 to 0.6. This can be done in 50 ml Falcon tubes. 3) Harvest the cells by centrifuging for 5 min at 1800 rpm. 4) Wash the cell pellet in 1.5 ml of ddH$_2$O by vortexing and transfer to a 2.0 ml Eppendorf. 5) Spin at highest speed for 5 s in a microcentrifuge. 6) Wash the cells in 1 ml TE/LiAc and spin again. 7) Resuspend the cells in 275 ml of TE/LiAc. 8) Add 30 ml of boiled ssDNA and 3-5 μg of the normalized AD-library. 9) Add 1.5 ml TE/LiAc/PEG and mix by inverting several times. 10) Incubate at 30° for 30 min to 1 h. 11) Heat shock at 42° for 15 min. 12) Spin down the cells. 13) Remove the supernatant and resuspend the cells in 900 ml of sterile ddH$_2$O. 14) Take 10 ml and add to 10 ml of ddH2O to create a 1:1000 dilution 15) To three 15 ml 3-AT plates, add approximately twenty-five 3 mm glass beads. Plate 300 μl of cells from step 12 on each 15 cm 3-AT plate. Spread the cells evenly by shaking the plates with glass beads.

Screening by mating

Y2H screening can be fermomed also by mating. DB and AD fusion proteins are transoformed in different mate type yeast strain: in this case DB in MaV203/Y8930 (*MATα*) and AD in MaV103/Y8800 (*MAT**a***). After mating, diployds cells can be screened in the same way of colonies resulting from a cDNA library screening [27].

Isolation of Two-Hybrid positives

1) After 4 to 5 days of growth, colonies with interacting proteins should have grown enough to isolate them. 2) Use a toothpick to pick colonies that grow above the background. Patch the colony in a small streak onto a 3-AT plate. 3) Incubate the plates at 30°C for 2 to 3 days. 4) Scrape a small clump of cells into 120 μl of SC-Leu-Trp medium in U-bottom plates. Seal the plates with airpore tape. 5) Incubate for 2 days at 30°C.

Phenotypic assays

1) Spot 5 μl of culture onto a SC-Leu-Trp plate. Grow at 30°C for 1 to 2 days. 2) Different plates to assay the two-hybrid reporters are needed: YPD for lacZ, 3-AT for HIS3, 5-FOA for URA3 (and -ADE for Y strains). YPD plate needs a nitrocellulose filter placed on it prior to replica plating. 3) Using replica velvets and a replica block, replica plate the yeast from the SC-Leu-Trp growth plate to the YPD/filter, 3-AT, and 5-FOA plates. Use the same velvet for each of the assay plates; there should be enough yeast on it for all three plates. 4) Replica clean the 3-AT and 5-FOA plates as follows. Use a clean velvet to remove excess yeast from the 3-AT plate. Repeat the procedure with the 5-FOA plate. 5) Culture all three assay plates at 30°C. After 1 day the YPD plate should have large spots of yeast on the filter. Perform β- Gal filter lift assays. 6). Examine 3-AT and 5-FOA plate. If the negative control shows growth after 1 day, replica clean again. 7) When the controls on the 3-AT and 5-FOA plates have grown to the appropriate levels remove plates and score results. 8) Score the 3-AT and 5-FOA plates along with the β-Gal filters. Any strain that passes at least two of the three tests is considered positive. Consolidate all of the positives into fresh plates. 9) Grow plates for 2 days and make glycerol stocks of the strains.

β-Gal filter lift assay

1) For each plate to be assayed, get one empty 1515 cm petri plate. Put two pieces of Whatman filter paper in the plate. Add 5 ml of β-Gal solution

to each plate. Let the paper soak up the solution remove bubbles under the Whatman paper. 2) Remove the nitrocellulose filter from the YPD (with yeast on the filter) and place in liquid nitrogen for at least 30 s. This lyses the cells. 3) Remove the filter from liquid nitrogen and allow it to thaw in air. Once the filter is flexible again, place in a petri plate with β-Gal solution-soaked Whatman paper. remove any bubbles that may be under the filter. 4) Put the β-Gal assay plates at 37°C overnight. The next day you can read the results with positives being blue; the stronger the positive, the stronger the blue [93].

Yeast PCR from yeast

It is necessary to isolate prey DNA from the yeast clones that pass the phenotypic assays and identify them by sequencing. Prey DNA can be obtained by polymerase chain reaction (PCR) using the universal primers on the vector. Primers to the activation domain sequence and the termination sequence are commonly used: activation domain CGCGTTTGGAATCAC-TACAGGG, termination sequence -GGAGACTTGACCAAACCTCTGGCG.

1) Spot 5 ml of the positives from the 96-well plate culture Culture at 30°C for 1 day. 2) Replica plate to YPD. Culture at 30°C for 1 day. 3) Add 15 ml of lysis solution to each well in a 96-well PCR plate. 4) Scrape some of the yeast cells off the YPD plate and resuspend into the 15 μl lysis solution. 5) Put the yeast at 37°C for 5 min then at 95°C for 5 min (this can be done in a PCR machine). 6) Set up the PCR plate using a final volume of 50 μl for each reaction. 7) Make a 1:10 dilution of the yeast lysis from step 8) Add 5 μl to the PCR plate. 9) Perform PCR using 5 min extension times.

PCR cycle is: 94°C, 4 min; (94°C, 45 sec; 56°C, 1 min; 68°C, 5 min)$\times$ 32 times; 72°C, 7 min; 4°C, indefinitely.

2.3.3 Cloning methods

A BP/LR Gateway (LR) recombination reactions strategy was that combines *att*B×*att*P and *att*L×*att*R reactions to transfer the PCR amplified Open Reading Frames (ORF) to a destinaction vector (pDEST-DB or pDEST-AD) starting from a donor vector (pDONR223) [113], [109]

Primer Design

The ORF must be PCR amplified using the following Gateway tails to subsequentely clone the ORF into the pDONR223:

- *att*B1.1: GGGGACAACTTTGTACAAAAAAGTTGGC

- *att*B2.1: GGGGACAACTTTGTACAAGAAAGTTGGGTA

The forward and reverse primers for PCR amplification are:

- 5' ORF Specific Primer (Forward primer): *att*B1.1 + 5'-3' ORF (usually 24 - 30 bp). You can omit the ATG start codon as the expression is regulated by the DB/AD fusion proteins in the destinaction vector.
- 3' ORF Specific Primer (Reverse primer): *att*B2.1 + reverse complement of the 3' terminus of the ORF (usually 24 - 30 bp) containing the termination codon.

Cloning ORF in the Donor vector: BP Reaction.

To clone the PCR amplified ORF into the donor vector pDONR223 assemble a BP reaction as follow:

*att*B1.1 PCR clone:	150 ng
pDONR223 vector:	150 ng
5× BP buffer:	2 μL
BP Clonase:	2 μL
TE bufer pH 8.0:	to 8 μL

Incubate at 25 °C for 1 hour to overnight. Add 2 μL of Proteinase K solution and incubate for 10 minutes at 37 °C. Transform into competent cells (e.g. DH5α) and plate on plates containing *Spectinomycin* 50 μg/mL.

The 5× BP buffer consists of: 100 mM Trsi-Cl (pH 7.5), 20 mM EDTA, 30 mM spermidine-HCl, 25% glycerol, 225 mM NaCl. Once you got you Entry clone, you can transfer the ORF into the destination vector.

Transferring the ORF into the Destinaction vector: LR Reaction.

To transfer ORF from the entry clone to the donor vector pDONR223 assemble an LR reaction as follow:

Entry clone:	50-150 ng
pDEST (AD or DB) vector:	150 ng
LR Clonase Enzyme Mix II:	2 μL
TE buffer pH 8.0 or H_2O:	to 8 μL

Alternatively you can use the following protocols:

Entry clone:	50-150 ng
pDEST (AD or DB) vector:	150 ng
LR Clonase Enzyme Mix II:	1 μL
(Total Volume 3 μL	

or, if using LR Clonase I

Entry clone:	100-150 ng
pDEST (AD or DB) vector:	75 ng
LR Clonase :	0.5 μL
5× LR buffer:	1 μL
TE buffer pH 7.5 :	to 5 μL

Incubate at 25 °C for 1 hour to overnight. Add 2 μL (or 1 μL if using the 5 μL volume reaction) of Proteinase K and incubate for 10 minutes at 37 °C. Transform into competent cells (e.g. DH5α) and plate on plates containing *Ampicillin* 100 μg/mL.

5× BP buffer Stock solution: for 10 M: 1 ml TRIS 1 M, 1 M EDTA 0.2 M, 1 M NaCl 2.25 M, 2.5 M of glycerol and 47 μL of spermidine (d=0.925 mg/ml). Finalize with MilliQ water and filter sterilize (0.22μm). Store at -20 °C.

Transformations of BP and LR products can be done in liquid cultures, with antibiotic selection of spectinomycin at 50 μg/mL (BP) or ampicillin at 100 μg/mL.

2.4 Testing two different yeast strains for Y2H experiments

2.4.1 Methodologies

Eight proteins out of a set of 54 different proteins previously utilized for both mating and cDNA library Y2H screening [81] were randonmly selected. The 8 open reading frames have been PCR amplified and subsequently cloned into destination vector pDEST-DB using a BP/LR Gateway recombination reactions strategy as previously described in section 2.3.3. To confirm that the design and the construction of the fusion proteins were correct, the vector/insert DNA junction have been sequenced to confirm the reading frame. The eight baits have been both transformed in the yeast strain Mav203 and Y8930 (Matα) following the transformation protocol as detailed in section 2.3.2.

The eight baits, transformed into MaV203, strain have been tested for auto-activation. There is indeed a major limitation inherent to the two-hybrid system. DB-X fusions can activate transcription independently of an interaction with an AD-Y protein (self activators) and thus cannot be used in conventional forward two-hybrid screens. Self activators include proteins that act as transcriptional activators in their respective organisms and maintain this ability in yeast or may also include They also include proteins that normally act in other processes but exhibit transcriptional activity when tethered to a promoter in yeast cells [46], [58]. Test for autoactivation was carried out by growing yeast on selective media -L+5FOA: self-activator-expressing colonies should not be able to grow under these conditions [143]. Two baits out of eight did show auto-activation behavior and were subsequently removed from screening.

The remaining six baits were then used to perform an adult human brain cDNA library (ProQuest, Invitrogen) screen as described in section 2.3.2. Bait strain containing single DB plasmid were individually transformed with 10 μg of the adult human brain cDNA library and 1-2 millions independent transformants have been on average screened for each bait using six 15 cm SC-L-T-H +3-AT plates. Baits transformed into MaV203 have been plated on 20 mM 3-AT plates while baits transformed in the new Y8930 strains have been plated on 0, 5, 10, 20 mM 3-AT plates to titrate the not yet fully characterized basal expression of the HIS3 gene in the Y8930. All primary positive clones were tested by further phenotypic assays using all three reporter genes: HIS3, URA3 and LacZ for MaV203 strain and HIS3, ADE2 and LacZ for Y8930 (see section 2.2.2). Phenotypic assays for HIS3, URA3 and ADE2 has been performed by growing colonies on media lacking the appropriate aminoacids as detailed in the Methods section.

Particular attention was paid in performing the β-Gal filter assay for positives arising from Y8930 transformed baits. With aim to optimize and adapt the protocol used for the MaV strain we performed β-Gal tests using different temperature conditions (room temperature and 37°C) and different periods of incubation, ranging from 15 min to overnight incubation.

All interactions were then retested by gap repair. The particular bait and prey proteins expressed in a yeast cell can have profound effects on yeast physiology, including changes in transcription rate, growth rate, viability, and cell permeability. Similar changes could occur due to mutations in the yeast genome arising in the course of a screen. When the mutations favor growth on the plates used to select positive interactions, false-positives arise. The gap-repair retest procedure eliminate all false-positives due to mutations in the yeast genome. This involves introducing into yeast the linearized pDEST-AD vector and the PCR-amplified insert AD-cDNA products into fresh yeast cells containing the bait plasmid. If there is some homology between the insert and the cut vector, the yeast will repair it by ligating the insert into the vector.

Retesting by gap repair is also necessary because the AD cDNA library plasmids does not carry the CYH2 negative selection marker that allows plasmid shuffling on cycloheximide-containing media [142]. Only the clones passing the gap repair retests were considered as true Y2H interactors.

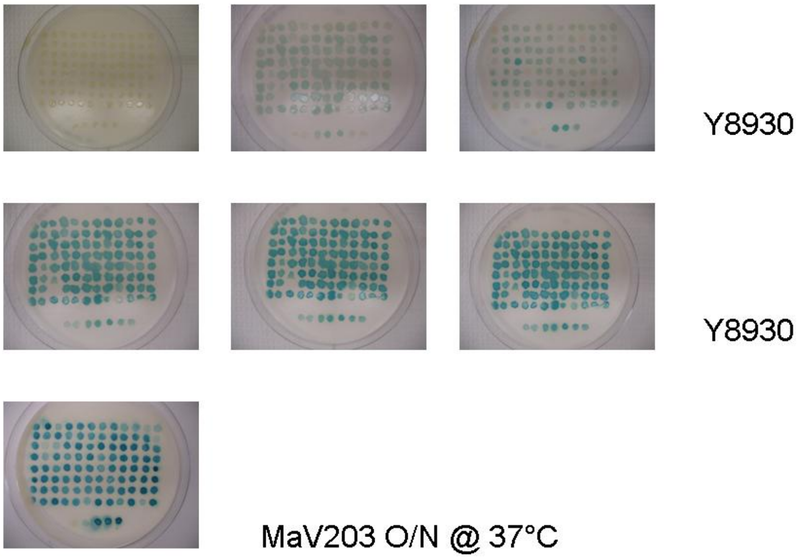


Figure 2.2: Comparison of **LacZ** for both Y8930 and MaV203 yeast strains. Six different incubations times (0, 30, 45, 60, 75, 120 min) at 37°C for Y8930 are shown together with an overnight incubation at 37°C for MaV203. For Y8930, fter 45-60 min blue levels are completely saturated.

2.4.2 Results and discussion

The comparison between the two cDNA library screening, performed using the same baits transformed in two different yeast strains was aimed to asses the advantages and disadvantages in the use of the two strains in performing Y2H experiments.

The analysis of the positives clones obtained by growing colonies on plates carrying different concentration of 3-AT showed that the basal expression of HIS3 in the Y8930 strains is several times lower than in MaV strain giving an almost undetectable background. Under these conditions, titration of the GAL1:HIS3 reporter for Y8930 seems not be anymore necessary. The reason for this is not known, since the HIS3 reporter locus is the same.

Both strains are very similar. Both are intended for *Gal4* based two-hybrid assays, and feature three reporter genes. Each uses *leu2* and *trp1* as auxotrophic markers to select for cells containing bait and prey plasmids. As a consequence, the same plasmids and libraries can be used in either strain.

Both use *HIS3* and LacZ reporter genes , but differ in the third reporter. MaV102 and MaV203 use *URA3*, which was added to confer sensitivity to 5-FOA. Instead of *URA3*, Y8800 and Y8930 use *ADE2*.

For the *HIS3* reporter, Y8800 and Y8930 need less 3-AT in the -His plates than MaV103 and MaV203 (0-5 mM compared to 20-100 mM). The reason for this is not known, since the *HIS3* reporter locus is the same.

The two strains showed a surprisingly different behavior in respect of the *LacZ* β-Gal filter assay. We did find that even incubation room temperature is sufficient to activate β-Galactosidase activity in Y8930. We did find that the typical turning to blue of this colorimetric testis is really faster in Y8930 than in MaV strain. If growth at 37°C, we did observed a full saturation of blue color for positive colonies and positive control, making impossible to separate not only weak interactors from strong ones but even positive from negative. Several incubation assays at room temperature did show that an optimum time to read the β-Gal filters ranges form 30 to 60 minutes while growing at 37°C dramatically decreases this time to few minutes. This behavior of the Y8930 strains make the *LacZ* β-Gal filter assay difficult to be performed not only if working in the frame of a high-throughput project.; it has been indeed difficult to optimize a well working protocol for *LacZ* test due to this tight time-scale even for small scale screenings. A comparison of β-Gal filter assay results for Y8930 (incubation at for different periods)is shown in Figure 2.2.

All preys previously found during the screening performed in [81] have been retrieved during our screen. The slightly highest number of preys detected using the Y8930 transformed baits can indicate a highest sensivity of the Y8930 strain in respect of the MaV strain to this kind of tests. Both *URA3* and *ADE2* yield less positives than the *HIS3* marker. This does not necessarily equate to being more stringent or identifying fewer false positives, as certain interactions can show stronger induction of *URA3* or *ADE2* than *HIS3*. For this is advisable to test all 3 reporter genes. Although the strains share the *HIS3* and *LacZ* reporters, the phenotypes resulting from them differ.

2.5 Conclusion

The two-hybrid system has some clear advantages over classical biochemical and genetic approaches. First of all it embodies an in vivo technique using the yeast host cell as a live test tube. This yeast system brings the higher eukaryotic reality closer than most in vitro approaches or techniques based on bacterial expression. Appealing features of this system are the minimal requirements to initiate a screening. Only the cDNA, full-length or even partial of the gene of interest is needed, in contrast to sometimes-high quantities of purified proteins or good quality antibodies needed in classical biochemical approaches.

Weak and transient interactions, often the most interesting in signaling cascades, are more readily detected in two-hybrid since the genetic reporter gene strategy results in a significant amplification. It is useful to keep in mind that there is a trade-off between the identification of weak interactions and the number of false positives encountered when performing a screening procedure. Apart from the ability to screen libraries, the two-hybrid system also allows for the analysis of known interactions. This can be achieved by pinpointing crucial residues for interaction or by a functional characterization of the entire subdomain. By doing semi-quantitative experiments one can even interpret affinities from two-hybrid experiments. It was demonstrated that the strength of interaction as predicted by the two-hybrid approach generally correlates with that determined in vitro, permitting discrimination of high-, intermediate- and low-affinity interactions [48]. In addition, binding affinities of peptides to retinoblastoma, as determined by surface plasmon resonance, correlated with results from the two-hybrid assay.

Anyway the most difficult steps in a Y2H screening are to determine if the interaction observed in yeast is direct and to place it in the right and meaningfull biological context. The tranactivation of the reporter constructed by the DB and AD fusion proteins indicate that these proteins formed a stable complex within the yeast nucleous; additonal component provided by the yeast itself could, anyway, contribute to this complex. To asses if the interaction is direct the bait and prey proteins can be synthesized as bacterial fusion proteins (i.e. to S-transferase and maltose binding proteins). These can be readily purfied and the interaction assessed *in vitro*. In case of interaction, the affinity can be determined.Affinity should be at least μM.

An interaction detected *via* yeast two-hybrid system may not be biologically relevant. For se exemple during a library screening, proteins form distinct cellular can be coexpressed, allowing interactions to take place that would normally not occur due to subcellular compartmentalizations. Further,proteins form different cells can be coexpressed, resulting in interaction that are not permitted in the organism. Finally, aberrant interactions may occur because of the two-hybrid proteins adopt nonphysiological configurations: this could arise because the hybrid proteins are truncated or incorrectly modified by the yeast or associated with yeast proteins.

The two-hybrid system was predicted to be limited to the analysis of cytoplasmic proteins. Indeed extracellular proteins or protein domains are often N-glycosylated and contain disulfide bonds, both of which are not expected to occur in the yeast nucleus [50]. However, several successes were reported with transmembrane receptors. Appropriate extracellular receptor-ligand interactions were demonstrated for the growth hormone,prolactin and growth hormone releasing receptors [148], []kajkowski1997igh. Thus, receptors with whole extracellular critical ligand binding determinants can sometimes be evaluated by the two-hybrid system. But, it may be inappropriate for receptors with determinants in transmembrane domains that form intramembraneous ligand binding pockets [65].

One of the most appealing features of the yeast two-hybrid system is that the identification of an interacting protein implies that at the same time the corresponding gene is cloned. Two-hybrid screens are sometimes referred to as functional screens, since interacting proteins might give a functional hint if at least one of the partners has a known functional commitment in a well understood signaling pathway. Trying to attribute function to an unknown target is often more difficult. Here, the identified partners need to be known or the problem will propagate. Although the outcome of a screening often results in many new hypotheses, they still need to be validated by other techniques.

Chapter 3

Evaluation of the quality of NMR protein structures

3.1 Introduction

3.1.1 Fundamentals of NMR theory

NMR spectroscopy and X-ray cristallography are currently the only techniques capable of determing the structures of biological macromolecules at atomic resolution. In addition, with NMR it is possile to study time dependent phenomena, such as intramolecular dynamics in macromolecules, reaction kinetics, molecular recognition or protein folding.

The limitations of NMR spectroscopy result from the low inherent sensitivity of the technique and from the high complexity and information content of NMR spectra of proteins. These problems are partially alleviated by new developments: the sensitivity and resolution of NMR are increased by progress in spectrometer technology. Progress in the theoretical and practical capabilities of this technique lead to a increasingly efficient utilization of the information content of NMR spectra.

Parallel developments in the biochemical methods (recombinant protein expression) allow the simple and fast preparation of protein samples. Heteronuclei like ^{15}N, ^{13}C and ^{2}H can be incorporated in proteins by uniform or selective isotopic labelling. Spectra from these samples can be drastically simplified. Additionally, some new informations about structure and dynamics of macromolecules can determined with these methods.

The phenomenon of magnetic resonance results from the interaction of the magnetic moment of an atomic nucleus μ with an external magnetic field. The cause of this magnetic moment is the quantum mechanical angular momentum (spin angular momentum) of all nuclei that have and odd number of protons and neutrons.

The spin is quantized according to

$$J = \hbar\sqrt{I(I+1)} \tag{3.1}$$

with J being the spin angular momentum, I the spin quantum number and $\hbar$ the reduced Planck's constant. The angular momentum and the magnetic moment are directly proportional:

$$\mu = \gamma J = \hbar\sqrt{I(I+1)} \tag{3.2}$$

The constant γ is the gyromagnetic ratio and is characteristic for each isotope. The sensitivity of a nucleus in NMR depends on γ: the higher the γ the higher the sensitivity is).

In an external magnetic field $\mathbf{B_0}$ the magnetic moment orients (if $\mathbf{B_0}= B_0 \vec{z}$ according to:

$$J_z = -m\hbar \tag{3.3}$$

$$\mu_z = \gamma m\hbar \tag{3.4}$$

The magnetic quantum number m can be an integer number between -*I* and +*I*. Thus, the external field leads to a splitting of the energy levels. For spin $\frac{1}{2}$ nuclei like protons, two energy levels exist according to a parallel or antiparallel orientation of the magnetic moment with respect to the magnetic field direction. The energy of these levels is given by the classical formula for a magnetic dipole in a homogenous magnetic field of the strength B_0:

$$E = -\mu_z B_0 \tag{3.5}$$

The magnetic moment of each nucleus precesses around $\mathbf{B_0}$. The frequency of this precession is the Larmor frequency ω_0 which is equivalent to the resonance frequency of the nucleus and the energy difference ΔE between the two levels.

$$\Delta E = \hbar\omega_0 \tag{3.6}$$

$$\omega_0 = \gamma B_0 \tag{3.7}$$

The Larmor frequency depends on the gyromagnetic ratio and the strength of the magnetic field and it is different for each isotope. At a magnetic field of 18.7 T the Larmor frequency of protons is 800 MHz.

A NMR sample contains many identical molecules (usually in a concentation range of mM for proteins). The spins of these molecules align indepently of each other parallel or antiparallel to the external field. The ratio of parallel spins to the antiparallel ones is given by the Boltzmann distribution:

$$\frac{N_p}{N_{ap}} = e^{-\frac{\Delta E}{kT}} \tag{3.8}$$

Both energy levels are nearly equally populated, because the energy difference is in the order of magnitude of thermic movements (kT). At a temperature of 300 K and a magnetic field of 18.7 T (800 MHz) the excess in the lower enery level is only 6.4 of 10000 particles for protons. This is the main reason for the inherently low sensitivity of NMR when compared to optical spectroscopic methods.

The magnetic moments of the individual spins sum up to a macroscopic magnetization M_0 which can estimated according to Curie's law:

$$M_0 = \gamma^2 N \hbar^2 \frac{1}{3kT} B_0 I(I+1) \tag{3.9}$$

It is the evolution of this macroscopic magnetization which is recorded in the spectrometer. The classical theory of NMR also deals with this quantity. In thermal equillibrium only magnetization along the axis of the magnetic field exists (by definition z), because the x and y components sum up to zero.

In the initial stage of investigation by NMR spectroscopy each resonance of the spectra must be associated to a specific nucleus in the investigated molecule. The aim of the analysis of NMR spectra is to extract all available information about interatomic distances and torsion angles. This process is called assignment.

The strategies employed for the assignment procedure depend on whether only homonuclear 2D spectra are available (unlabelled proteins), whether ^{15}N heteronuclear spectra are available (^{15}N labelled proteins) or whether triple resonance spectra (^{15}N/^{13}C doubly labelled proteins) are available, but in general, the assignment can be divided in two parts.

The sequential assignment of the amino acids in the protein sequence and the assignment of the amino acid side chains. Experiments as 2D COSY and TOCSY are employed for the identification of amino acid spin systems. The 2D NOESY experiment is used to sequentially connect the spin systems. The fist step in sequential assignment is the identification of certain amino acids in COSY/TOCSY, with a characteristic pattern of cross peaks like glycine, alanine, threonine, valine, leucine and isoleucine. Glycine contains two H$^{\alpha}$ protons and is therefore readily identified. Valine, leucine and isoleucine can be recognized by their two methyl groups which give a characteristic row of double signals between 0 and 1.5 ppm. In the same way, alanine and threonine are identified by their single methyl groups.

In the second stage of the assignment process, the sequential contacts from the already identified amino acids to the neighboring ones are searched for in the 2D NOESY spectra. The connectivity of a given amino acid in the sequence i to its following one $i+1$ can be observed in the NOESY because the distance of the amide proton of $i+1$ to the i-th H^{α}, H^{β} or H^{δ} protons of i-th is usually smaller than 5 Å. Therefore, sequential cross signals to H(i), H^{β}(i) etc. are observed at the frequency of $\mathrm{H}^{N}(i+1)$. These interresidual signals can be distinguished from the intraresidual ones by comparing the 2D NOESY with the 2D TOCSY spectrum. A series of these sequential peaks between $\mathrm{H}^{\alpha}(i)$ and $\mathrm{H}^{N}(i+1)$ determines the order $(i, i+1, i+2, ...)$ of the amino acid spin systems.

Thus, dipeptides are identified and subsequently prolonged to oligopeptides by the search for further sequential contacts. Some time along the line these oligopeptides can be placed at a unique place in the primary structure by comparison with the amino acid sequence of the protein - they are sequentially assigned.

The chain of sequential connectivites is interrupted by proline residues because these have no amide proton. Therefore, no H^{N}(i)-H^{α}(i-1) cross signal can be observed. However, if the proline (i) is in its trans conformation, the sequential $\mathrm{H}^{N}(i-1)$-$\mathrm{H}^{\delta}(i-1)$ and $\mathrm{H}^{\alpha}(i-1)$-$\mathrm{H}^{\delta}(i)$ cross signals can be observed.

Another problem is, that this approach of sequential assignment breaks down for larger proteins because the vast number of signals leads to spectral overlap which hinders the identification of signals.

So far, the emphasis has been on identification of the observed signals in the spectra and their correlation with the amino acid protons giving rise to the signals. Afterwards, one has to extract the data which are relevant for the structure. Of special importance in this respect are proton-proton distances, which can be estimated from the signal intensities in the different spectra .

Signal intensity depends on the distance r between two nuclei i and j, according to [1] :

$$NOE_{ij} \propto \frac{1}{r_{ij}^6} \tag{3.13}$$

The distances are derived from the spectra after calibration against NOE signals for known distances (such as distances in elements of secondary structure) and grouped into a few classes. An upper and a lower bound of distance is assigned to each class. The lower bound is often set to the sum of the van der Waals radii of the two protons. In this procedure, all non-sequential signals which are visible in the NOESY spectra have to be assigned, the number of which easily exceeds 1000 in a medium-sized protein (ca. 120 amino acids). It is distinguished between cross peaks of protons no more than five amino acids apart in the protein sequence (medium range NOEs) and those which are more than five amino acids apart (long range NOEs). The former are mainly indicative of the protein backbone conformation and are used for secondary structure determination, whereas the latter are an expression of the global structure of the protein and therefore contain the main information used for tertiary structure calculation.

3.1.2 NMR spectroscopy does allow for major errors

Ca. 14% of the protein structures deposited in the Protein Data Bank (PDB) [15] have been solved by NMR. NMR has also been recognized as an important component in Structural Genomics projects worldwide, as it can contribute significantly to their success rate (in terms of structures solved with respect to the number of targets entering the pipeline).

[1] The actual equations to determine NOEs size and sign for a steady state (the NOE $\eta_{I(J)}$ is observed on signal I when J is saturated) are

$$\eta_{I(J)} = \frac{\sigma_{I(J)}}{\rho_{I(J)}} \tag{3.10}$$

with

$$\sigma_{I(J)} = \left(\frac{\mu_0}{4\pi}\right)^2 \frac{2\hbar^2\gamma_I^2\gamma_J^2 J(J+1)}{15r_{IJ}^6}\left[\frac{6\tau_c}{1+(\omega_I+\omega_J)^2\tau_c^2} - \frac{\tau_c}{1+(\omega_I-\omega_J)^2\tau_c^2}\right] \tag{3.11}$$

$$\rho_{I(J)} = \left(\frac{\mu_0}{4\pi}\right)^2 \frac{2\hbar^2\gamma_I^2\gamma_J^2 J(J+1)}{15r_{IJ}^6}\left[\frac{6\tau_c}{1+(\omega_I+\omega_J)^2\tau_c^2} + \frac{3\tau_c}{1+\omega_I^2\tau_c^2} + \frac{\tau_c}{1+(\omega_I-\omega_J)^2\tau_c^2}\right] \tag{3.12}$$

where μ_0 is the permeability of the vacuum and τ_c is the correlation time.

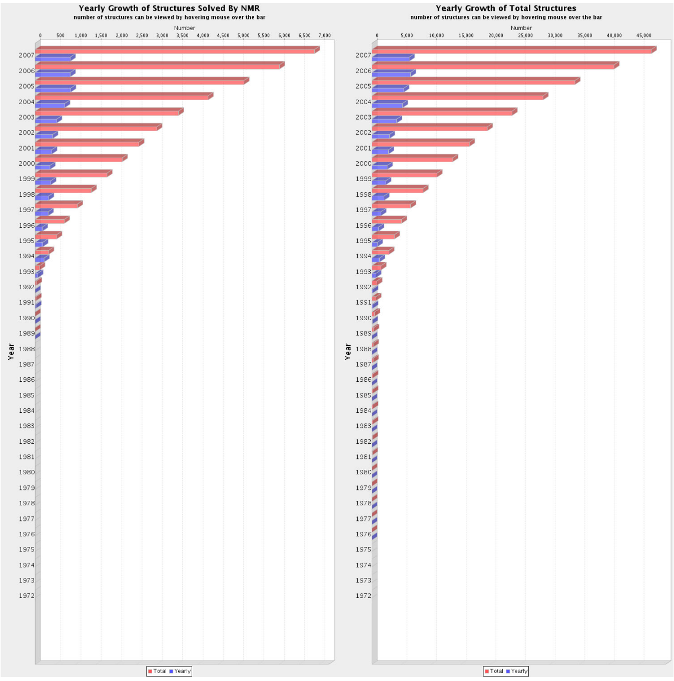


Figure 3.1: Yearly growth report of structure deposited within the PDB as November 2007. On the left is shown the yearly growth of NMR solved structures, on the right the yearly growth of total structures. PDB counts a grand total of 47136 structures deposited. The 92% are protein structures, the remaining are nucleid acids (NA) or NA-protein complexes (4% respectively). The 14% of structures (6880) is solved by NMR. Source: `www.wwpdb.org`.

NMR structures are complementary to X-ray structures under many aspects. When structure determination is coupled with, as often happens, characterization of protein backbone dynamics, one gains information beyond the single snapshot of a static structure and actually obtains direct experimental insight into protein flexibility. Flexible regions are often handled with difficulty in NMR structure determination, because of the paucity (or even absence) of restraints that high mobility causes. These regions therefore are (very) poorly defined, their conformation often being driven by the force field used in calculation more than by experimental data. This may yield a significant decrease of overall accuracy. On the other hand, NMR is particularly important to characterize systems that are partly or mostly unstructured. These systems contain structurally ordered regions next to highly mobile regions, which prevent their crystallization. When compared with X-ray crystallography, NMR is particularly suited to investigate complexes with dissociation constants sub-micromolar or higher, which are quite dynamic, and whose crystallization is difficult and may easily "freeze" configurations that are not the most relevant to extract biological insights. This is one of the present frontiers of NMR. How these systems should be handled, and in particular whether coordinates for the structure of ordered regions or of all the protein should be released, is a completely open issue. At present, and in the next future, NMR is getting more and more important in addressing intermolecular contacts, also thanks to experimental strategies that allow signals to be detected even in systems of very high molecular mass.

A proposal for standardization of the presentation of NMR structures has been put forward in 1998. Since then, validation and quality assessment of NMR structures has progressed significantly.

In NMR structure determination, there exist only nascent methods for the statistically rigorous estimation of precision, and active research is ongoing in this area. [124]. Traditionally, NMR structures have been deposited in the Protein Data Bank (PDB) [15] as ensembles which are the result of replicating the structure determination procedure several times. The conformational variability across this ensemble can then be used as a measure of reproducibility, and therefore the precision. Measurement of the divergence of an ensemble of structures typically involves calculating a superposition, minimizing the root-mean-square deviation (RMSD, see section 3.2.1, relation (3.14) and following) from an average or representative set of atomic coordinates. This approach suffers from pitfalls, however, in that the set of atoms being superimposed must be well-chosen [125]. Furthermore, this strategy also implicitly assumes that the ensemble deposited in the PDB characterizes well the uncertainty inherent in the data [126], [98], [110]. The estimation of accuracy also has complicating factors, not the least of which is the fact that the correct structure is typically unknown. In the absence of the knowledge of the true structure, one can attempt to estimate accuracy using internal measures. This is now routinely done in macromolecular X-ray crystallography using the free R-factor [32], which provides a cross-validated estimate of the goodness-of-fit that is correlated with the phase error. The direct application of such a strategy in NMR structure determination is difficult because of the combined effect of low data density and high information content of individual data points [95] (e.g. critical long-range nuclear Overhauser effect (NOE) restraints), but extensions based on a jackknife approach have been suggested [24]. More sophisticated approaches based on the assessment of the degree to which the model satisfies the raw data have also been described [23], [59].

Throughout the years several errors have been uncovered in the PDB, which often resulted in the replacement of the incorrect models with improved ones. Systematic errors and mistakes often result from errors in the interpretation of the experimental data and relate directly to the accuracy of the final structure models. For example, in NMR spectroscopy errors can be introduced by misassignment of the spectral signals. Many different types of errors can be identified in protein structures, ranging from too tightly restrained bond lengths and angles, to molecules exhibiting a completely incorrect fold. The first type of errors often does not have large consequences for the analysis of the structure and typically can be easily remedied by refinement in a proper force field. It as been shown indeed that before refinement, side-chain planarity is often too tightly or too loosely restrained, with a local minimum there where X-ray structures are commonly found. After refinement NMR structures usually have planarity RMSD comparable to high-resolution X-ray structures, shwoing much better agreement with presently acceptable simple stereochemistry parameters such as bond lengths and bond angles [83] [96]. For structures determined using NMR spectroscopy there are cases where re-evaluation of the experimental data has resulted in the replacement of structures in the PDB. For example, in the structure of the oligomerization domain of p53 a difference in the orientation of the two dimers was observed between the NMR [37] and crystal structure that was released shortly after the NMR structure [63] (PDB entry 1C26).

Re-examination of the nuclear Overhauser enhancement (NOE) data led to the identification of three misinterpreted peaks in the original p53 NOE assignments and the inclusion of several new NOEs, resulted in a revision of the original PDB entry [38] (PDB entry 1OLH). Similar is the case of the anti-σ factor AsiA [71] a low number of misinterpreted NOE signals (17 in total) resulted in a largely incorrect fold (the now obsolete PDB entry 1KA3). In this case, it was not until a second solution structure of AsiA was published [139] (PDB entry 1JR5) that the experimental data of the original AsiA structure were reexamined and the assignment errors were discovered [72].

Stimulated by the activities of the Coordination Action "NMR-Life" and by a round table on protein structure quality and related problems helded in the frame of the 11th Chianti Workshop on Magnetic Resonance, we thought a critical and systematic re-examination of majors validation tools was needed.

Program	**Accessible at** http:\\	**Reference**
FRST	`protein.cribi.unipd.it\frst`	Tosatto 2005
HARMONY	`caps.ncbs.res.in\harmony`	Pougalenthi et al. 2006
HOPscore	`hoppscore.lbl.gov\run.html`	Sims and Kim 2006
MolProbity	`molprobity.biochem.duke.edu`	Davis et al. 2007
PROCHECK	`www-nmr.cabm.rutgers.edu\PSVS`	Laskowski et al. 1993, 1996
ProsSA	`prosa.services.came.sbg.ac.at\prosa.php`^{a}	Wiederstein and Sippl 2007
TAPscore	`protein.cribi.unipd.it\tap`	Tosatto and Battistutta 2007
WHAT CHECK	`swift.cmbi.kun.nl\WIWWWI`	Hooft et al. 1996

Table 3.1: List of analysed tools. Web addresses are shown together with references. ^{a}This server is actually accessible at `https:\\`.

3.2 Methodologies

3.2.1 Tools to assess protein structures quality

A wide range of computational quality parameters have been developed and reviewed over the years, using different criteria and different strategies. In the past, claims have been even raised about the fact that the majority of validation tools tends to overestimate the number of errors because the fact of the high number of false positive tough theseclaims have been disproved [91], [74].

Generally speaking, it is possible to distinguish geometric, energetic and conformational criteria. Geometric criteria are mainly standard values for bond lengths and angles derived from small molecule data. [39] [57] These form strong restraints and are generally enforced during the refinement process, so they possess little validation power. Energetic criteria are based on evaluation of interaction preferences or profiles [86], [146] [92]. These methods can provide insight into the quality of the structure but their interpretation in experimental terms and feedback into the refinement process is rather difficult.

The most promising validation criteria are based on conformational criteria. The best example is the Ramachandran plot [107] of backbone (ϕ, ψ) torsion angles. While each amino acid type may, in theory, adopt a large number of different conformations, large areas of the Ramachandran plot are almost empty. This is due to steric clashes deriving from the local geometry of the polypeptide chain. The main chain (ϕ, ψ) torsion angles are usually not restrained during refinement and this makes the Ramachandran plot a powerful validation tool. These parameters alone are however, as mentioned, insufficient by themselves to unambiguously identify "bad" structures, as advanced structure calculation protocols including structural refinement in explicit solvent (although always recommendable) may often result in good scores even for inaccurate structures. Many of these scores are also poorly useful when calculations have been performed including database potentials such as Ramachandran plot refinement [68]. Several tools have been developed to estimate the quality of a protein model based on the Ramachandran plot. HOPPscore has been recently developed to take into account higher order backbone torsion angle maps [121].

Of these, PROCHECK [75] and WHAT CHECK [57] are perhaps the most frequently used methods for validation in NMR spectroscopy WHAT IF currently is the most used, if not the most complete tool, for assessing the quality of a structure and checking errors. It has become is a de facto standard for structure and structural model validation. as they are used for judging structures to be deposited in the PDB, combining several stereochemical checks and measures of torsion angle compatibility. Several of these methods (e.g. WHAT CHECK) are able to pinpoint the really wrong structures through a detailed analysis of different aspects of protein structures. Once a structure falls into the range of roughly plausible folds however the situation becomes more complicated. It is possible to construct structures with acceptable values for the standard criteria that are largely incompatible with the protein sequence. This means that given a roughly plausible structures, it is almost possible to quantify the degree of "nativeness" and highlight the best structures.

The community of bio-NMR scientists that are engaged in the structural characterization of proteins is still lacking a generally accepted measure of accuracy of the structures produced. Although several measures have been proposed over the years, and also in the recent past, each laboratory is still using its own ensemble of quality checks and parameters. Even when different laboratories are using the same measures, they often compute them differently or assess these measures using subjective, non-reproducible criteria. Therefore, it is often the case that when analysed by different experts, the same structure will be assessed through different statistics, and thus assigned a different quality level.

Commonly used measures of accuracy are the size and number of residual restraint violations and the statistics of the distribution of residues in the regions of the Ramachandran plot. A related issue is that of the measure of precision of an NMR structure, which is typically evaluated by the Root Mean Square Deviation (RMSD) of the backbone atoms or of all heavy atoms within the bundle of conformers that is released in the PDB. RMSD is a measure of the average distance between the backbones of superimposed structures. Given two sets of point s (i.e. two structures S_2 and S_2), RMSD Ξ between

S_1 and S_2 is given by

$$\Xi = \sqrt{\frac{1}{N}\sum_i (s_{1i} - s_{2i})^2} \tag{3.14}$$

Given a family F of K structures S, the RMSD within the family can be defined as

$$\Xi(F) = \frac{1}{K(K-1)}\sum_{i>j} \Xi(S_i, S_j) \tag{3.15}$$

or

$$\Xi(F) = \frac{1}{K}\sum_k \sqrt{\frac{1}{N}\sum_i (s_i - \bar{s_i})} \tag{3.16}$$

where $\bar{s_i}$ indicates the $i-th$ point of the mean structure over the family. The equation (3.15) gives the RMSD as the average value of the RMSDs between all the structures in the family. (3.16) gives the RMSD as the average value of RMSD of each structure in respect to the averaged structure of the family. However, these measures often leads to an overestimation of a structure's precision.

All the above kind of measures have been shown in the literature to be inadequate to properly function as indicators of structure quality. Moreover, they fail to clearly distinguish correct from wrongly folded structures. In fact, solution structures can be refined to very good values for these parameters, and thus apparently show high accuracy and precision, while still having major flaws. Although the two aspects of precision and accuracy of NMR structures are clearly linked to one another, we will focus mainly on the aspects related to accuracy. Note that both are affected by the criteria used for inclusion of a given conformer in the final bundle of structures (e.g. global energy *versus* sum of restraint violations) in an unpredictable way.

Accuracy can be assessed at different levels. One aspect is global fold accuracy, i.e. whether the NMR structure really reproduces the protein fold. At the other extreme, another aspect of accuracy is local conformation, e.g. of side chains, which can have a deep impact also on important features such as hydrogen bonding or protein core compactness. Global accuracy is typically assessed by means of global statistics, which range from scores describing the total number of violations (sometimes normalized by the total number of restraints used) or the residual largest violations, to more refined measures such as RDC-based Q-factors (see section 3.3 for a brief overview), but can also include measures such as Z-scores which is the deviation of

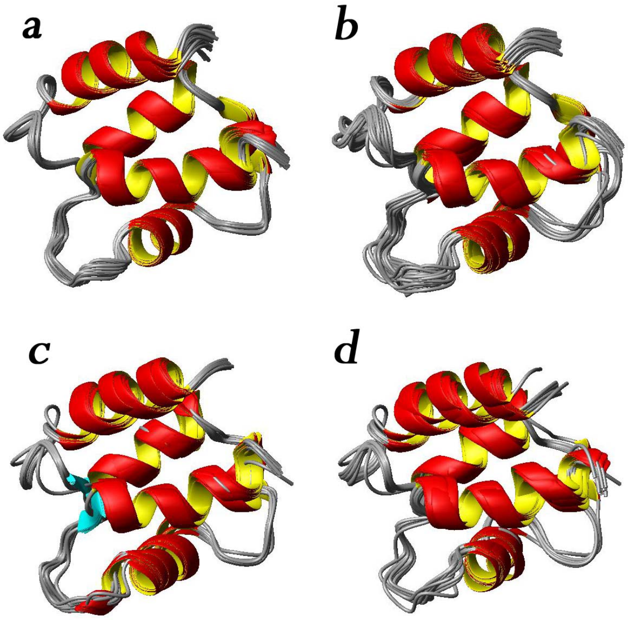


Figure 3.2: The deposited reference structures of Calbindin D_{9K} is shown (panel *a*) together with the corresponding 25, 50 and 75% perturbed structure families (panels *b*, *c* and *d* respectively.)

that quality indicator from a database-derived average value, in units of the standard deviation of that database derived average. The database derived average will by definition have a Z-score of 0. Z-scores can be calculated with different definitions by a variety of programs.

3.2.2 Validation of validation tools

Without sake of completeness we reviewed several tools, choosen among the most popular or the more recent released. They represent the state of the art of programs and softwares dedicate to the assessment of the quality of three-dimensional structure of proteins. The following programs were used: FRST [136], HARMONY [106], HOPPscore [121], MolProbity [42], PROCHECK [75], ProSA (ProSA-Web) [146], TAP [135], WHAT IF (WHAT CHECK) [57]. All these tools are freely available through web servers. A list of web addresses is shown in table Table 3.1. Details about these programs can be found in the original publications or briefly summarized in [115].

Among the methods analyzed ProSA (ProSA-web) TAP, FSRT and HARMONY try to give an unique parameter indicating the overall quality of a

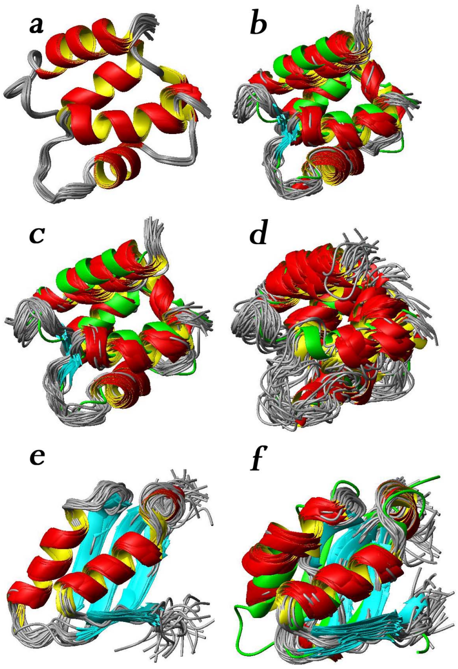


Figure 3.3: The three flawed families calculated for Calbindin are shown in panels *b*, *c* and *d*. The reference 1KQV structure is shown in *a*. The flawed family calculated for MNK6 is shown in *f* whereas the reference family 1YJR is shown in sub-plot *e*. In panels *b*-*d* and *f*, the first model of the reference family is shown in green superimposed to the flawed bundle.

structural model, while other tools give more extensive outputs listing several different parameters usually concerning conformational and geometrical aspects. We can refer to programs like ProSA and TAP/FSRT as statistical methods for the reason that the scores spring from the mean values of forces or potentials. In this way they give a parameter accounting for the overall quality of a model and also an estimate of the local quality, usually given as a per-residue plot of the energy of pseudo energy. The availability of a unique parameter to fully describe the goodness of a structure is an appealing idea especially for the occasional consumer not familiar with concepts like rotamer normality or C^{β} deviations and just looking, for instance, for a good template for homology modelling.

Structure	**RMSD (Å)**	**ref. RMSD (Å)**
PDZ (1Y7N)	0.68	-
PDZ recalc	0.81	0.60
PDZ perturbed 25%	0.63	1.23
PDZ perturbed 50%	0.36	1.44
PDZ perturbed 75%	0.38	1.67
Calb D9k (1KQV)	0.33	-
Calb D9k recalc	0.71	0.67
Calb D9k perturbed 25%	0.50	1.34
Calb D9k perturbed 50%	0.28	1.50
Calb D9k perturbed 75%	0.25	1.79
Calb D9k Error 1	0.82	2.65
Calb D9k Error 2	0.79	2.67
Calb D9k Error 3	2.23	2.94
Menkes 6 (1YJR)	1.07	-
Menkes 6 Error	1.32	3.48

Table 3.2: Backbone RMSDs values for the various structures. The RMSD column contains RMSD values within the bundle of conformers. The RMSD to the reference is the RMSD of the flawed structure with respect to the first model of the reference structure (equation (3.15).

3.2.3 A proteins benchmark for tools assessment

The selected tools have been tested and validated on several protein test systems in which errors have been introduced. The selected test proteins were: bovine Calbindin D_{9K}, a vitamin D-dependent calcium-binding protein, the second PDZ domain of the human neuronal adaptor X11α and the apo form of the A69P mutant of the sixth soluble domain of the Menkes protein. All deposited models have been retrieved from the RCSB Protein Data Bank (PDB).

Calbindin D_{9K} (UniProt code P02633) is a 75 aminoacid calcium-binding protein, whose structure consists of 4 helices and 3 loops. A total of 1675 meaningful NOE data, 37 dihedral angles and 1097 pseudocontact shifts have been used for structure calculations. The PDB deposited ensemble 1KQV [17] of structures has been used as reference. We selected this protein because the fact of the great number of experimental data available; it has been indeed extensively used as a test protein for a variety of different studies [16], [12], [20].

The second PDZ domain of the human neuronal adaptor X11α (UniProt code Q02410, PDB entry 1Y7N, [47]) is a 90 aminoacid residue protein and its fold consists of 2 helices and 6 beta sheets. This is a high quality NMR structure determined by a group deeply involved in the development of structure validation and refinement methodologies. We may therefore consider it as a state of the art NMR-solved proteins. The original experimental data (1725 meaningful NOE's and 95 dihedral angle restraints) have been retrieved from the BioMagResBank [138] (mrblock_id 51978, bmrb_id 6113).

The structure of MNK6 (UniProt code Q04656) has been recalculated using 1956 meaningful NOE data and 82 dihedrals angles. MNK6 is a 75 residues protein with a $\beta\alpha\beta\beta\alpha\beta$ fold (PDB entry 1YJR [9]). As long as the fold of this proteins is correct, some regions are not structurally perfectly assessed due the lack of NOE assignment in loops regions. This makes this protein an interesting benchmark to prove the discriminatory power and sensivity of validation tools.

We did generated flawed structures still having a plausible fold mimicking unaccurated NMR experimental data The original experimental distance constraints have been randomly perturbed by adding and/or subtracting up to 75% of their original values. These data have been subsequently used to perform structure calculations. More severe errors were generated introducing wrong NOE assignments in the data set of Calbindin D_{9K} and the Menkes. In the first case we switched the NOE constraints for the protons of residues Lys7 and Lys41. In the second case, in addition to the previous misassignments, we interchanged all NOE values of Lys25 and Lys16. In the third case we switched all the assignments of residues Glu51 and Glu64. Similarly, for the sixth domain of the Menkes protein we switched all the NOE assignments of residues Cys18 and Cys35. It should be noticed that these residues have been chosen randomly and not because they play a particular or critical role in protein folding or functionality. Some RMSD data for the recalculated families are summarized in Table 3.2. The effects of this errors and inaccuracies are shown in Figure 3.2 and Figure 3.3.

3.2.4 Structure calculations

Calculation of the 3D structure is formulated in the program CYANA as a minimization problem for a target function that measures the agreement between a structure and the given set of constraints. The CYANA target function [53], [54] is given by the equation

$$V = \sum_{c=u,l,v} w_c \sum_{(\alpha,\beta)\in I_c} (d_{\alpha\beta} - b_{\alpha\beta})^2 + w_a \sum_{(\alpha,\beta)\in I_a} \left[1 - \frac{1}{2} \left(\frac{\Delta_i}{\Gamma_i} \right)^2 \right] \Delta_i^2 \qquad (3.17)$$

where α and β are two atoms,and $b_{\alpha\beta}$ the upper and lower bounds on the distances $d_{\alpha\beta}$ between the two atoms. $\Gamma_i = \pi - (\theta_i^{max} - \theta_i^{min})/2$ denotes the half width of the forbidden range of the i-th torsion angle value θ_i that is allowed in the interval $[\theta_i^{min},\ \theta_i^{max}]$. I_u, I_l, I_v are the sets of atom pairs (α, β) with upper, lower or van der Waals distance bounds respectively. I_a is the set of restrained torsional angles. The parameters w_u, w_l, w_v, w_a are the weighting factors of the different types of constraints. Δ_i is the size of the i-th torsion angle constraint violation.

The CYANA target function V (equation (3.17)) is zero if and only if all experimental distance constraints and torsion angles constraints are fulfilled and all non-bonded pairs satisfy a check for the absence of steric overlap. A conformation that satisfies the constraints more closely than another one will have a lower target function value.

The minimization algorithm is based on simulated annealing [67] by molecular dynamics simulation in torsion-angle space. Molecular dynamics simulation includes kinetic energy that allows overcoming barriers of the potential surface which reduce the problems of trapping in local minima. Molecular dynamics simulation using torsion angles instead of cartesian coordinates as degree of freedom [54], [127], [69], [127], [90] provides a very efficient way to calculate NMR structures. The only degree of freedom are the torsion angles that is the rotations about the single bonds, such as the conformation of the molecule is uniquely specified by the values of all torsion angles. Covalent bonds that are incompatible with a tree structures are treated by distance constraints.

The potential energy landscape of a protein surface is complex and studded with many local minima. The temperature schedule is important because the fact that kinetic energy determines the maximal height of energy barriers that can be overcome in a molecular dynamics trajectory. A standard simulated annealing protocol consist of five steps:

1. *Initial minimization.* 100 conjugate gradient minimization step are performed including only distance constraints up to 3 residues apart along the sequence followed by a 100 minimization steps including all constraints is performed to reduce high-energy interactions. All hydrogen atoms are excluded form the check for steric overlap.

2. *High-temperature phase.* A torsion angle dynamics at constant high temperature. One-fifth of all N torsion angle dynamics steps are performed at a constant high reference temperature (typically 10^4 K). The time step is initialized to 10^{-15} s.

3. *Slow cooling.* The remaining $\frac{4}{5}N$ are performed during which the tem-

perature approaches to zero according a fourth-power law.

4. *Low-temperature phase with individual hydrogen atoms.* All hydrogen atoms are incorporated into the check for steric overlap. 100 conjugate gradient minimization steps are performed, followed by 200 torsion angle dynamics steps at zero reference temperature.

5. *Final minimization.* A final minimization of 1000 conjugate gradient steps.

3.2.5 Paramagnetic based constraints

The presence of a paramagnetic metal ion induces a shift on the nuclear resonances. This shift is determined by two contributions: A contact contribution, due to through-bond nuclear spin electron spin coupling, and a pseudocontact shift contribution. The pseudocontact term is due to the dipolar interaction between a nuclear magnetic moment and an average induced electron magnetic moment. The latter depends on the scalar product of the metal magnetic susceptibility tensor with the applied magnetic field vector. As a result, the pseudocontact shift values depend on the position of each observed nucleus in the magnetic metal susceptibility frame, with origin on the metal ion, and on the anisotropy of the latter, according to the following equation [17]

$$\delta_i^{PCS} = \frac{1}{12\pi r_{iM}^3}\left[\Delta\chi_{ax}(3\cos^2\theta_i - 1) + \frac{3}{2}\Delta\chi_{rh}\sin^2\theta_i\cos 2\phi_i\right] \tag{3.18}$$

where r_{iM} is the distance between the nucleus i and the metal ion M; θ_i and ϕ_i are the polar angles of the vector r_{iM} with respect to the principal axes of the magnetic susceptibility tensor χ centered on the metal ion. The axial and rhombic magnetic susceptibility anisotropies are given by

$$\Delta\chi_{ax} = \chi_{zz} - \frac{\chi_{xx} + \chi_{yy}}{2} \tag{3.19}$$

$$\Delta\chi_{rh} = \chi_{xx} - \chi_{yy} \tag{3.20}$$

The amount of information provided by these restraints can be so significant that diamagnetic proteins containing a metal binding site may be conveniently investigated by substituting the diamagnetic metal ion with a paramagnetic one [20]. It may be convenient to substitute different paramagnetic metal ions in the same binding site, in order to have several sets of data, which are often complementary [19]. Indeed, the metal susceptibility tensor depends on the nature and the coordination properties of the metal ion and therefore different metal ions provide independent information. PCS shifts have been sucessfully implemented in CYANA package [11]. They can be used together with NOEs constraints for structure calculations. A new term V^{PCS} has been added to the CYANA target function V (equation (3.17)):

$$V^{PCS} = \sum_i w_i \left[\max(|\delta^{PCS}_{i_{calc}} - \delta^{PCS}_{i_{obs}}| - T_i, 0)\right]^2 \tag{3.21}$$

where w_i is the weight of the i-th constraint and T_i is the tolerance on the i-th proton; it gives an extimation of the error with respect to the shift real value.

3.2.6 Structure refinement

Rapid calculation of protein structures necessitates severe simplifications of the nonbonded interactions. The unrealistic treatment of electrostatic and van der Waals interactions can lead to nonoptimal packing, unsatisfied hydrogen bond donors or acceptors. For these structures, validation programs often yield low quality indices such as the Ramachandran plot appearance or the side-chain packing quality. Refinement in a full molecular dynamics force field is one way to improve the quality of the structures [83]. The program AMBER 8 that implements the SANDER module [104] has been used to perform energy minimization of each structure within the family generated during the molecular dynamic calculation with CYANA.

SANDER use first order *steepest descent* and *conjugate gradient* algorithms for energy minimization. The force field contains bond terms (bendings, stretchings, torsions) and non bond terms (Van der waals, hydrogens and electrostatic interactions). The potential function U is given by:

$$\begin{aligned} U = \quad & \sum_{bonds} K_r(r-r_0)^2 + \sum_{bonds} K_\theta(\theta-\theta_0)^2 + \sum_{bonds}\sum_n \frac{V_n}{2}\left[\cos(\eta_n - \gamma_n)\right] + \\ + \quad & \sum_{i<j} \epsilon_{ij}\left[\left(\frac{R_{ij}}{r_{ij}}\right)^{12} - \left(\frac{R_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{i<j}\frac{q_i q_j}{r_{ij}} \end{aligned} \tag{3.22}$$

where r is the actual distance between two bonded atoms, r_0 the equilibrium distance, K_r the stretching force costant, θ the actual bond angle, θ_0 the equilibrium angle and K_θ an associated costant. V_n is the potential of the torsional barrier, γ_n the pahse of the dihedral angle, η_n the barrier periodicity, q_i and q_j the partial charges of the the i-th and j-th atom respectively at distance r_{ij}; R_{ij} is the sum over the Van der wallas radii R_i and R_j.

A potential function N^{NOE}, weighted on NOEs values, is also implemented to take account of these restraints during the minimization. It is a mixed linear-armonic potential in the form (distances in Å):

$$N^{NOE} = \begin{cases} k_1(d_{ij} - d_{ij}^u)^2 & \text{for} \quad d_{ij}^u < d_{ij} < d_{ij}^u + 0.5 \\ k_1(2.0 - d_{ij})^2 & 1.5 < d_{ij} < 2.0 \\ 0 & 2.0 < d_{ij} < 2.0 \\ k_2(d_{ij} - d_{ij}^u)^2 & d_{ij} > d_{ij}^u + 0.5 \\ k_2(1.75 - d_{ij})^2 & d_{ij} < 1.5 \end{cases} \tag{3.23}$$

where d_{ij}^u is the upper limit for the distance d_{ij} between atoms i and j. To the costant k_1 and k_2 have been assigned values of $133.8\,\mathrm{kJmol}^{-1}\text{Å}^{-2}$ and $133.8\,\mathrm{kJmol}^{-1}\text{Å}^{-1}$ respectively.

Pseudocontact shifts have been introduced in the potential function by mean of the term

$$V^{PCS} = \sum_i K_{PCS} \left[\max(|\delta_{i_{calc}}^{PCS} - \delta_{i_{obs}}^{PCS}| - T_i, 0)\right]^2 \tag{3.24}$$

where K_{PCS} is force constant. The sum runs over all the i shifts considered [11].

The minimization procedure is stopped when the difference of the energy or of the norm of the energy gradient do not vary between two subsequent minimization steps in respect to a fixed threshold. The threshold is setted to $10^{-7}\,\mathrm{kcalmol}^{-1}$ and $0.1\,\mathrm{kcalmol}^{-1}\text{Å}^{-1}$ respectively.

3.3 Discussion

Over the last years fine ideas and tools for assessment of the quality or validation of NMR structures have been proposed by several groups working on this topic, and more results will be produced in the future. In this work we gave a first evaluations of the relative merits of different measures or strategies implemented through several different programs (see section 3.2.1).

We think that it is worthy to notice that, in general, authors make efforts to prove the performance of their tools showing how many times the proposed methods is able to discriminate a flawed structure from a good one, sometimes working on huge dataset analysing thousand of proteins or models; on the contrary the problem of false positive is rarely (if never) discussed. In the past claims about false-positives generated by validation tools have been raised and discussed [91]. Anyway we think that false positives are not a real problem in the sense that a false positive usually should lead to double check experimental data and the structure: it is up to the "structure producer" and to his/her experience to judge if and when data or structure need reexamination and or correction. It is obvious, that producers should resist the temptation to refine their models against the output of a certain validation tool, i.e adjusting the Ramachandran plot appearance for, as we showed, this is not a guarantee of a correct structure. This behaviour also induces a dangerous negative feedback as in the case of X-ray structures, usually used to build reference databases for geometrical and stereochemistry properties or in the building of test dataset used to develop and asses new validation tools.

From our findings, we can point out that false negative are much more worrisome: when a structure is flagged as problematic, it is likely it is. When a structure is flagged as a good structure, this can not be trusted completely.

The kind of errors introduced may be considered relatively gross. Nevertheless, because they lead to significant distortions of the structures while maintaining several features intact, such as individual secondary structure elements, they are representative of perhaps more subtle errors that can arise e.g. due to misinterpretation of experimental data. Indeed, the kind of errors discussed here, that preserve a roughly plausible fold, leads to a relatively satisfactory evaluation of the tools that assess the global quality of the structure most likely because the fold is not incorrect. However, many aspects of the structure are wrong, and may become misleading when one attempts to use the structure to interpret biochemical/biological evidence.

We did compare structures with correct structure to structures deriving from an artificially wrong NMR input. We can remark that often but not always, the scores for the correct structures are better than for the wrong ones as in the case of TAP score. But this is not sufficient for a validation tool; having a structure set with a TAPscore of 0.6866-0.7143 (numbers from Table 3 of the attached paper), how to know that a better one would have a TAPscore of 0.7576-0.7946? In a real case one has only one set and comparisons are not possible. The many inconsistencies revealed during our screening suggest that it is not possible to use of these tools for the systematic screen of large numbers of models without any visual inspection of the structures. This is particularly true especially for those tools whose output is an unique parameter indicating the overall quality of a structural model. The spread in score over the members of the bundle of unperturbed conformers is relatively large and makes these tools almost useless in discriminating between correct and flawed structures. Score ranges typically allowed for NMR structures comprise both good and flawed structures without consistent differentiation. Indeed, significantly perturbed structures can bypass several

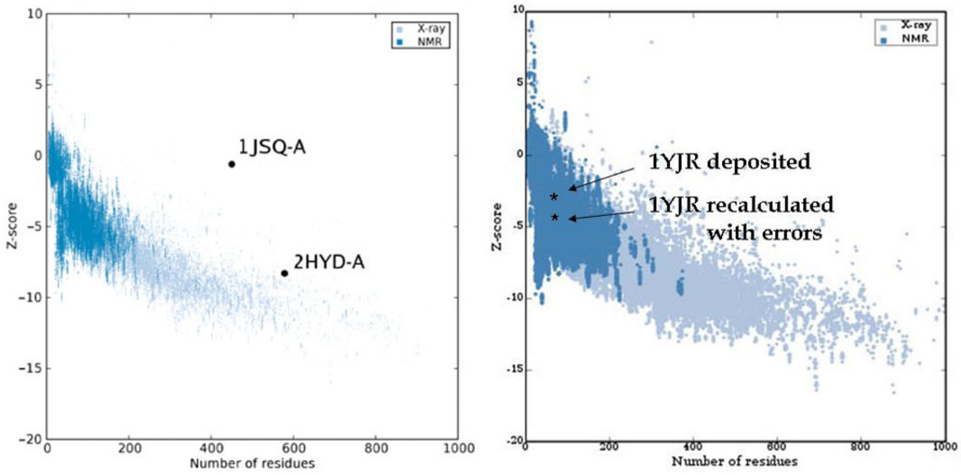
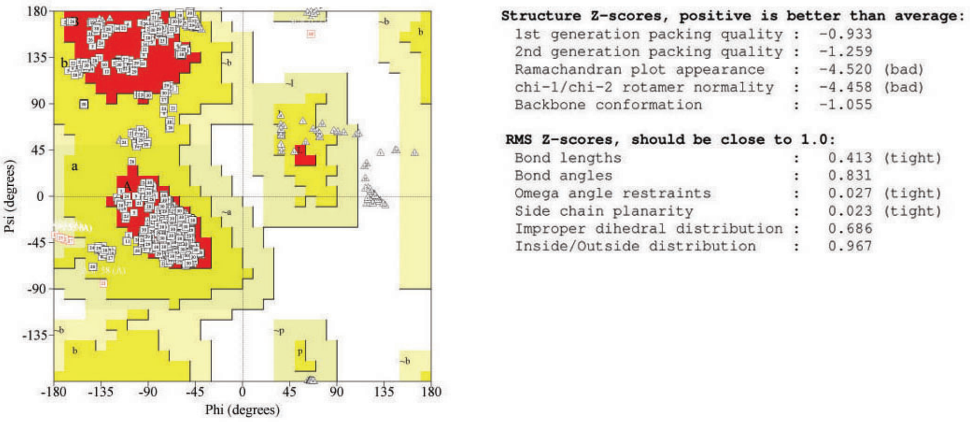


Figure 3.4: Left: Investigation of two ABC transporter structures using the ProSA-web service. The structure of a monomer of MsbA (PDB code 1JSQ, chain A [30]) determined by X-ray crystallography to 4.5 Å resolution and retracted due to problems in the interpretation of the crystallographic raw data. Although homologous to 1JSQ, the structure of the monomer of Sav1866 (PDB code 2HYD, chain A [43]) as determined by X-ray crystallography to 3.0 Å resolution differs considerably from the 1JSQ A chain The ProSA (ProSA-web) Z-scores indicate that 2HYD has features characteristic for native structures. This is not true in the case of the flawed recalculated structure of 1YJR (Right). In this case ProSA Z-scores fail to discriminate between wrong and correct structurec assigning a lower (-4.78) (that is more "native") score to the wrong model in respect to the correct one (-3.72). This is due to the score spread within the bundle.

validation tools, also due to the relatively high intra-family variation that we observed. In Figure 3.4 shows the effect of score spread within the bundle in the case of a ProSA analysis of the recalculated Menkes protein (1YJR): the best wrong model scores -4.78 against the -3.72 of the worse correct model, thus assigning a more plausible fold to the flawed structure.

Operatively, it appears that the distribution of (ϕ, ψ) torsion angles pairs, either in the "simple" Ramachandran plot (taking into account only the percentage of residues in the core region) or in the more sophisticated analysis performed by HOPPscore, remains one the best indicators of structure quality, provided that backbone dihedral angle restraints are applied loosely as done here. Clearly, refinement of the structure against a potential based on the Ramachandran map would make this indicator useless. WHAT CHECK



```
Structure Z-scores, positive is better than average:
  1st generation packing quality :  -0.933
  2nd generation packing quality :  -1.259
  Ramachandran plot appearance   :  -4.520 (bad)
  chi-1/chi-2 rotamer normality  :  -4.458 (bad)
  Backbone conformation          :  -1.055

 RMS Z-scores, should be close to 1.0:
  Bond lengths                   :   0.413 (tight)
  Bond angles                    :   0.831
  Omega angle restraints         :   0.027 (tight)
  Side chain planarity           :   0.023 (tight)
  Improper dihedral distribution :   0.686
  Inside/Outside distribution    :   0.967
```

Figure 3.5: Ramachandran plot (from PROCHECK) for the deposited 1KQV family of Calbindin D_{9K} paired with the correspondent WHAT IF checks report.

Z-scores for $\chi_1 - \chi_2$ rotamer normality or for backbone conformation are also informative. Note however that the $\chi_1 - \chi_2$ rotamer distribution can be significantly affected by the force field applied in molecular dynamics calculations/refinement, as these dihedrals may often not be well determined by the NMR data. Incidentally we did notice that the spread of scores and the backbone RMSD within the bundle do not correlate.

It is worthy to be noticed that none of the evaluated tools checks the consistency of the structure with the corresponding experimental NOEs. The selected test proteins are all compact and of similar size. For the largest one (PDZ domain, 90 residues) there are 1725 NOEs, while for the smaller ones (Calbindin D_{9K}, Menkes sixth domain, 75 residues) there are 1675 or even 1956 NOEs. The number of not 'meaningful' NOEs may bias some of the statistics, e.g. average restraint violation. Short distances in the structure should correspond to NOEs and vice versa. This needs some parameters (cutoff for short distances), and one may argue also about other aspects such as dynamics. Similar problems exist for most scores; consistency between the list of NOEs and the resulting structure would be a validation directly on the experimental input, and a rather informative way for NMR spectroscopists.

This could be accomplished with some useful tools have been proposed recently [95] to identify within the several hundreds of restraints typically used in an NMR structure calculation those that exert the greatest influence on the resulting structure (these "unique" restraints are identified using information analysis as those having the highest information content among all restraints used), which are thus worth checking most carefully. This is a local-level analysis that however may impact crucially on the global fold.

Ideas based on information retrieval analysis have been exploited to propose NMR quality factors based on Recall, Precision (not to be confused with structural precision; here it refers to completeness of back-calculated NOESY peak lists relative to experimental peak list data) and performance. A normalized F-measure statistic, dubbed discriminating power (DP), has been also proposed [59] as a measure of the capability of the experimental data to discriminate between a given 3D structure and a freely rotating chain. These parameters provide global measures of the goodness-of-fit of the 3D structures with NOESY peak lists, and thus do not directly assess the accuracy of the structure.

It should be kept in mind that structure determination strategies are typically iterative and comprise various steps of structure analysis/error correction/addition of new NOEs, in which the only goal is to minimize the RMSD of the bundle and the residual violations of upper distance limits. Care should be taken to evaluate quality parameters also during the iterative calculation process and not just at the end, in order to detect potential problems as early as possible.

A protocol on how to use existing tools to check an/or validate calcultated structures prior to deposition in the PDB that has received general approval by the bulk of the community is still lacking. A possible first step is the compilation of a list of checks/quality factors to be always performed at the end of a solution structure determination project, and to be released at the time of structure deposition or together with the publication describing the structure. "Instructions" on how these checks should be performed in a standardized manner should be drawn up as well.

It is likely that this approach would result in tighter and better ranges of quality scores for deposited NMR structures.

Chapter 4

Publications

4.1 Monomorphysm of human cytochrome *c*

4.2 The War of Tools: how can NMR spectroscopists detect errors in their structures?

Bibliography

[1] A. A. Marchler-Bauer and S. H. Bryant. A measure of progress in fold recognition? *Proteins: Struct. Funct. Genet*, 37:218–225, 1999.

[2] E. AB, A. R. Atkinson, L. Banci, I. Bertini, S. Ciofi-Baffoni, K. Brunner, T. Diercks, V. Dótsch, F. Engelke, G. E. Folkers, C. Griesinger, W. Gronwald, U. Gúnther, M. Habeck, R. N. de Jong, H. R. Kalbitzer, B. Kieffer, B. R. Leeflang, S. Loss, C. Luchinat, T. Marquardsen, D. Moskau, K.-P. Neidig, M. Nilges, M. Piccioli, R. Pierattelli, W. Rieping, T. Schippmann, H. Schwalbe, G. Travé, J. Trenner, J. Wóhnert, M. Zweckstetter, and R. Kaptein. NMR in the SPINE Structural Proteomics project. *Acta Crystallographica Section D*, 62(10):1150–1161, Oct 2006.

[3] M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, and R. F. Moreno. Complementary dna sequencing: expressed sequence tags and human genome project. *Science*, 252(5013):1651–1656, Jun 1991.

[4] S. Albeck, P. Alzari, C. Andreini, L. Banci, I. M. Berry, I. Bertini, C. Cambillau, B. Canard, L. Carter, S. X. Cohen, J. M. Diprose, O. Dym, R. M. Esnouf, C. Felder, F. Ferron, F. Guillemot, R. Hamer, M. Ben Jelloul, R. A. Laskowski, T. Laurent, S. Longhi, R. Lopez, C. Luchinat, H. Malet, T. Mochel, R. J. Morris, L. Moulinier, T. Oinn, A. Pajon, Y. Peleg, A. Perrakis, O. Poch, J. Prilusky, A. Rachedi, R. Ripp, A. Rosato, I. Silman, D. I. Stuart, J. L. Sussman, J.-C. Thierry, J. D. Thompson, J. M. Thornton, T. Unger, B. Vaughan, W. Vranken, J. D. Watson, G. Whamond, and K. Henrick. SPINE bioinformatics and data-management aspects of high-throughput structural biology. *Acta Crystallographica Section D*, 62(10):1184–1195, Oct 2006.

[5] T. Alber, D. P. Sun, J. A. Nye, D. C. Muchmore, and B. W. Matthews. Temperature-sensitive mutations of bacteriophage t4 lysozyme occur

at sites with low mobility and low solvent accessibility in the folded protein. *Biochemistry*, 26(13):3754–3758, Jun 1987.

[6] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

[7] S. Amari, F. Beltrame, J. Bjaalie, T. Dalkara, E. De Schutter, G. Egan, N. Goddard, C. Gonzalez, S. Grillner, A. Herz, et al. Neuroinformatics: the integration of shared databases and tools towards integrative neuroscience. *Journal of Integrative Neuroscience*, 1(2):117–128, 2002.

[8] M. Bahar, C. Ballard, S. X. Cohen, K. D. Cowtan, E. J. Dodson, P. Emsley, R. M. Esnouf, R. Keegan, V. Lamzin, G. Langer, V. Levdikov, F. Long, C. Meier, A. Muller, G. N. Murshudov, A. Perrakis, C. Siebold, N. Stein, M. G. W. Turkenburg, A. A. Vagin, M. Winn, G. Winter, and K. S. Wilson. SPINE workshop on automated X-ray analysis: a progress report. *Acta Crystallographica Section D*, 62(10):1170–1183, Oct 2006.

[9] L. Banci, I. Bertini, F. Cantini, M. Migliardi, A. Rosato, and S. Wang. An atomic-level investigation of the disease-causing a629p mutant of the menkes protein, atp7a. *J Mol Biol*, 352(2):409–417, Sep 2005.

[10] L. Banci, I. Bertini, A. Rosato, and G. Varani. Mitochondrial cytochromes c: a comparative analysis. *J Biol Inorg Chem*, 4(6):824–837, Dec 1999.

[11] L. Banci, I. Bertini, G. G. Savellini, A. Romagnoli, P. Turano, M. A. Cremonini, C. Luchinat, and H. B. Gray. Pseudocontact shifts as constraints for energy minimization and molecular dynamics calculations on solution structures of paramagnetic metalloproteins. *Proteins*, 29(1):68–76, Sep 1997.

[12] R. Barbieri, I. Bertini, G. Cavallaro, Y.-M. Lee, C. Luchinat, and A. Rosato. Paramagnetically induced residual dipolar couplings for solution structure determination of lanthanide binding proteins. *Journal of the American Chemical Society*, 124(19):5581–5587, 2002.

[13] P. L. Bartel, J. A. Roecklein, D. SenGupta, and S. Fields. A protein linkage map of escherichia coli bacteriophage t7. *Nat Genet*, 12(1):72–77, Jan 1996.

[14] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. Genbank. *Nucleic Acids Res*, 35(Database issue):21–25, Jan 2007.

[15] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res*, 28(1):235–242, Jan 2000.

[16] I. Bertini, G. Cavallaro, M. Cosenza, R. Kümmerle, C. Luchinat, M. Piccioli, and L. Poggi. Cross correlation rates between curie spin and dipole-dipole relaxation in paramagnetic proteins: the case of cerium substituted calbindin d9k. *J Biomol NMR*, 23(2):115–125, Jun 2002.

[17] I. Bertini, A. Donaire, B. Jiménez, C. Luchinat, G. Parigi, M. Piccioli, and L. Poggi. Paramagnetism-based versus classical constraints: an analysis of the solution structure of ca ln calbindin d9k. *J Biomol NMR*, 21(2):85–98, Oct 2001.

[18] I. Bertini, E. Grassi, C. Luchinat, A. Quattrone, and E. Saccenti. Monomorphism of human cytochrome c. *Genomics*, 88(5):669–672, Nov 2006.

[19] I. Bertini, M. B. Janik, Y. M. Lee, C. Luchinat, and A. Rosato. Magnetic susceptibility tensor anisotropies for a lanthanide ion series in a fixed protein matrix. *J Am Chem Soc*, 123(18):4181–4188, May 2001.

[20] I. Bertini, M. B. Janik, G. Liu, C. Luchinat, and A. Rosato. Solution structure calculations through self-orientation in a magnetic field of a cerium(iii) substituted calcium-binding protein. *J Magn Reson*, 148(1):23–30, Jan 2001.

[21] M. S. Boguski, T. M. Lowe, and C. M. Tolstoshev. dbest–database for "expressed sequence tags". *Nat Genet*, 4(4):332–333, Aug 1993.

[22] J. U. Bowie, J. F. Reidhaar-Olson, W. A. Lim, and R. T. Sauer. Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science*, 247(4948):1306–1310, Mar 1990.

[23] A. T. Brünger, G. M. Clore, A. M. Gronenborn, R. Saffrich, and M. Nilges. Assessing the quality of solution nuclear magnetic resonance structures by complete cross-validation. *Science*, 261(5119):328–331, Jul 1993.

[24] Brunger, Axel T., Clore, G. Marius, Gronenborn, Angela M., Saffrich, Rainer, and Nilges, Michael. Assessing the quality of solution nuclear magnetic resonance structures by complete cross-validation. *Science*, 261(5119):328–331, jul 1993.

[25] V. Brusic and N. Petrovsky. Immunoinformatics: the new kid in town. *Novartis Found Symp*, 254:3–13, 2003.

[26] K. H. Buetow, M. N. Edmonson, and A. B. Cassidy. Reliable identification of large numbers of candidate snps from public est data. *Nat Genet*, 21(3):323–325, Mar 1999.

[27] G. Cagney, P. Uetz, and S. Fields. High-throughput screening for protein-protein interactions using two-hybrid assay. *Methods Enzymol*, 328:3–14, 2000.

[28] M. A. Calderwood, K. Venkatesan, L. Xing, M. R. Chase, A. Vazquez, A. M. Holthaus, A. E. Ewence, N. Li, T. Hirozane-Kishikawa, D. E. Hill, M. Vidal, E. Kieff, and E. Johannsen. Epstein-Barr virus and virus human protein interaction maps. *Proceedings of the National Academy of Sciences*, 104(18):7606–7611, 2007.

[29] M. Cargill, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, N. Shaw, C. R. Lane, E. P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G. Q. Daley, and E. S. Lander. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet*, 22(3):231–238, Jul 1999.

[30] G. Chang and C. B. Roth. Structure of msba from e. coli: a homolog of the multidrug resistance atp binding cassette (abc) transporters. *Science*, 293(5536):1793–1800, Sep 2001.

[31] D. Chasman and R. M. Adams. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J Mol Biol*, 307(2):683–706, Mar 2001.

[32] Z. Chen, E. Blanc, and M. S. Chapman. Improved free r factors for cross-validation of macromolecular structure - importance for real-space refinement. *Acta Crystallogr D Biol Crystallogr*, 55(Pt 1):219–224, Jan 1999.

[33] P. Chevray and D. Nathans. Protein Interaction Cloning in Yeast: Idetification of Mammalian Proteins that React with the Leucine Zipper

of Jun. *Proceedings of the National Academy of Sciences*, 89(13):5789–5793, 1992.

[34] A. M. Chinnaiyan, K. O'Rourke, B. R. Lane, and V. M. Dixit. Interaction of ced-4 with ced-3 and ced-9: a molecular framework for cell death. *Science*, 275(5303):1122–1126, Feb 1997.

[35] K. Y. Choi, B. Satterberg, D. M. Lyons, and E. A. Elion. Ste5 tethers multiple protein kinases in the map kinase cascade required for mating in *S. cerevisiae*. *Cell*, 78(3):499–512, Aug 1994.

[36] A. Christoffels, A. van Gelder, G. Greyling, R. Miller, T. Hide, and W. Hide. Stack: Sequence tag alignment and consensus knowledgebase. *Nucleic Acids Res*, 29(1):234–238, Jan 2001.

[37] G. M. Clore, J. G. Omichinski, K. Sakaguchi, N. Zambrano, H. Sakamoto, E. Appella, and A. M. Gronenborn. High-resolution structure of the oligomerization domain of p53 by multidimensional nmr. *Science*, 265(5170):386–391, Jul 1994.

[38] G. M. Clore, J. G. Omichinski, K. Sakaguchi, N. Zambrano, H. Sakamoto, E. Appella, and A. M. Gronenborn. Interhelical angles in the solution structure of the oligomerization domain of p53: correction. *Science*, 267(5203):1515–1516, Mar 1995.

[39] C. Colovos and T. O. Yeates. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci*, 2(9):1511–1519, Sep 1993.

[40] J. A. Daniel, J. Yoo, B. T. Bettinger, D. C. Amberg, and D. J. Burke. Eliminating Gene Conversion Improves High-Throughput Genetics in Saccharomyces cerevisiae. *Genetics*, 172(1):709–711, 2006.

[41] S. Dao-pin, D. E. Anderson, W. A. Baase, F. W. Dahlquist, and B. W. Matthews. Structural and thermodynamic consequences of burying a charged residue within the hydrophobic core of t4 lysozyme. *Biochemistry*, 30(49):11521–11529, Dec 1991.

[42] I. W. Davis, L. W. Murray, J. S. Richardson, and D. C. Richardson. Molprobity: structure validation and all-atom contact analysis for nucleic acids and their complexes. *Nucleic Acids Res*, 32(Web Server issue):615–619, Jul 2004.

[43] R. J. Dawson and K. P. Locher. Structure of a bacterial multidrug abc transporter. *Nature*, 443(7108):180–185, Sep 2006.

[44] C. Diaz-Camino, E. P. Risseeuw, E. Liu, and W. L. Crosby. A high-throughput system for two-hybrid screening based on growth curve analysis in microtiter plates. *Anal Biochem*, 316(2):171–174, May 2003.

[45] C. M. Drysdale, D. W. McGraw, C. B. Stack, J. C. Stephens, R. S. Judson, K. Nandabalan, K. Arnold, G. Ruano, and S. B. Liggett. Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc Natl Acad Sci U S A*, 97(19):10483–10488, Sep 2000.

[46] W. Du, M. Vidal, J. E. Xie, and N. Dyson. Rbf, a novel rb-related gene that regulates e2f activity and interacts with cyclin e in drosophila. *Genes Dev*, 10(10):1206–1218, May 1996.

[47] A. E. Duquesne, M. Ruijter, J. Brouwer, J. W. Drijfhout, S. B. Nabuurs, C. A. Spronk, G. W. Vuister, M. Ubbink, and G. W. Canters. Solution structure of the second pdz domain of the neuronal adaptor x11alpha and its interaction with the c-terminal peptide of the human copper chaperone for superoxide dismutase. *J Biomol NMR*, 32(3):209–218, Jul 2005.

[48] J. Estojak, R. Brent, and E. Golemis. Correlation of two-hybrid affinity data with in vitro measurements. *Molecular and Cellular Biology*, 15(10):5820–5829, 1995.

[49] S. Field and O. Song. A novel genetic system to detect protein protein interactions. *Nature*, 340(6230):245–246, 1989.

[50] S. Fields and R. Sternglanz. The two-hybrid system: an assay for protein-protein interactions. *Trends Genet*, 10(8):286–92, 1994.

[51] M. J. Fogg, P. Alzari, M. Bahar, I. Bertini, J.-M. Betton, W. P. Burmeister, C. Cambillau, B. Canard, M. Carrondo, M. Coll, S. Daenke, O. Dym, M.-P. Egloff, F. J. Enguita, A. Geerlof, A. Haouz, T. A. Jones, Q. Ma, S. N. Manicka, M. Migliardi, P. Nordlund, R. Owens, Y. Peleg, G. Schneider, R. Schnell, D. I. Stuart, N. Tarbouriech, T. Unge, A. J. Wilkinson, M. Wilmanns, K. S. Wilson, O. Zimhony, and J. M. Grimes. Application of the use of high-throughput technologies to the determination of protein structures of bacterial and viral pathogens. *Acta Crystallographica Section D*, 62(10):1196–1207, Oct 2006.

[52] N. Guex, A. Diemand, and M. C. Peitsch. Protein modelling for all. *Trends Biochem Sci*, 24(9):364–367, Sep 1999.

[53] P. Güntert, W. Braun, and K. Wüthrich. Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program diana and the supporting programs caliba, habas and glomsa. *J Mol Biol*, 217(3):517–530, Feb 1991.

[54] P. Güntert, C. Mumenthaler, and K. Wüthrich. Torsion angle dynamics for nmr structure calculation with the new program dyana. *J Mol Biol*, 273(1):283–298, Oct 1997.

[55] M. K. Halushka, J. B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz, and A. Chakravarti. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat Genet*, 22(3):239–247, Jul 1999.

[56] L. Holm and C. Sander. Mapping the protein universe. *Science*, 273(5275):595–603, Aug 1996.

[57] R. W. Hooft, G. Vriend, C. Sander, and E. E. Abola. Errors in protein structures. *Nature*, 381(6580):272–272, May 1996.

[58] G. Hu, S. Zhang, M. Vidal, J. L. Baer, T. Xu, and E. R. Fearon. Mammalian homologs of seven in absentia regulate dcc via the ubiquitin-proteasome pathway. *Genes Dev*, 11(20):2701–2714, Oct 1997.

[59] Y. Huang, R. Powers, and G. Montelione. Protein NMR Recall, Precision, and F-measure Scores (RPF Scores): Structure Quality Assessment Measures Based on Information Retrieval Statistics. *Biophys. Chem*, 4:51–120, 1994.

[60] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P. S. Langendijk-Genevaux, M. Pagni, and C. J. Sigrist. The prosite database. *Nucleic Acids Res*, 34(Database issue):227–230, Jan 2006.

[61] K. Irizarry, V. Kustanovich, C. Li, N. Brown, S. Nelson, W. Wong, and C. J. Lee. Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nat Genet*, 26(2):233–236, Oct 2000.

[62] P. James, J. Halladay, and E. A. Craig. Genomic Libraries and a Host Strain Designed for Highly Efficient Two-Hybrid Selection in Yeast. *Genetics*, 144(4):1425–1436, 1996.

[63] P. D. Jeffrey, S. Gorina, and N. P. Pavletich. Crystal structure of the tetramerization domain of the p53 tumor suppressor at 1.7 angstroms. *Science*, 267(5203):1498–1502, Mar 1995.

[64] W. Y. Jeng, J. H. Shiu, Y. H. Tsai, and W. J. Chuang. Solution structure of reduced recombinant human cytochrome c. *To be published.*

[65] E. Kajkowski, L. Price, M. Pausch, K. Young, and B. Ozenberger. Investigation of growth hormone releasing hormone receptor structure and activity using yeast expression technologies. *J Recept Signal Transduct Res*, 17(1-3):293–303, 1997.

[66] P. Kellam and M. Alba. Virus bioinformatics: databases and recent applications. *Appl Bioinformatics*, 1(1):37–42, 2002.

[67] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. Optimization by simulated annealing. *Science*, 220(4598):671–680, may 1983.

[68] G. J. Kleywegt and T. A. Jones. Phi/psi-chology: Ramachandran revisited. *Structure*, 4(12):1395–1400, Dec 1996.

[69] G. R. Kneller and K. Hinsen. Generalized euler equations for linked rigid bodies. *Phys. Rev. E*, 50(2):1559–1564, Aug 1994.

[70] R. Koch, H. G. van Luenen, M. van der Horst, K. L. Thijssen, and R. H. Plasterk. Single nucleotide polymorphisms in wild isolates of caenorhabditis elegans. *Genome Res*, 10(11):1690–1696, Nov 2000.

[71] L. J. Lambert, V. Schirf, B. Demeler, M. Cadene, and M. H. Werner. Flipping a genetic switch by subunit exchange. *EMBO J*, 20(24):7149–7159, Dec 2001.

[72] L. J. Lambert, V. Schirf, B. Demeler, M. Cadene, and M. H. Werner. Flipping a genetic switch by subunit exchange. *EMBO J*, 23(24):7149–7159, Dec 2004.

[73] E. S. Lander. The new genomics: Global views of biology. *Science*, 274(5287):536–539, oct 1996.

[74] R. A. Laskowski, M. W. MacArthur, and J. M. Thornton. Validation of protein models derived from experiment. *Curr Opin Struct Biol*, 8(5):631–639, Oct 1998.

[75] R. A. Laskowski, J. A. Rullmannn, M. W. MacArthur, R. Kaptein, and J. M. Thornton. Aqua and procheck-nmr: programs for checking the quality of protein structures solved by nmr. *J Biomol NMR*, 8(4):477–486, Dec 1996.

[76] I. Letunic, R. R. Copley, S. Schmidt, F. D. Ciccarelli, T. Doerks, J. Schultz, C. P. Ponting, and P. Bork. Smart 4.0: towards genomic data integration. *Nucleic Acids Res*, 32(Database issue):142–144, Jan 2004.

[77] A. Levo, A. Koski, I. Ojanperä, E. Vuori, and A. Sajantila. Post-mortem snp analysis of cyp2d6 gene reveals correlation between genotype and opioid drug (tramadol) metabolite ratios in blood. *Forensic Sci Int*, 135(1):9–15, Jul 2003.

[78] S. Li, C. M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P. O. Vidalain, J. D. Han, A. Chesneau, T. Hao, D. S. Goldberg, N. Li, M. Martinez, J. F. Rual, P. Lamesch, L. Xu, M. Tewari, S. L. Wong, L. V. Zhang, G. F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H. W. Gabel, A. Elewa, B. Baumgartner, D. J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S. E. Mango, W. M. Saxton, S. Strome, S. Van Den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K. C. Gunsalus, J. W. Harper, M. E. Cusick, F. P. Roth, D. E. Hill, and M. Vidal. A map of the interactome network of the metazoan c. elegans. *Science*, 303(5657):540–543, Jan 2004.

[79] S. B. Liggett. Molecular and genetic basis of beta2-adrenergic receptor function. *J Allergy Clin Immunol*, 104(2 Pt 2):42–46, Aug 1999.

[80] S. B. Liggett, L. E. Wagoner, L. L. Craft, R. W. Hornung, B. D. Hoit, T. C. McIntosh, and R. A. Walsh. The ile164 beta2-adrenergic receptor polymorphism adversely affects the outcome of congestive heart failure. *J Clin Invest*, 102(8):1534–1539, Oct 1998.

[81] J. Lim, T. Hao, C. Shaw, A. J. Patel, G. Szabó, J. F. Rual, C. J. Fisk, N. Li, A. Smolyar, D. E. Hill, A. L. Barabási, M. Vidal, and H. Y. Zoghbi. A protein-protein interaction network for human inherited ataxias and disorders of purkinje cell degeneration. *Cell*, 125(4):801–814, May 2006.

[82] J. J. Lima, M. H. Mohamed, T. H. Self, L. V. Eberle, and J. A. Johnson. Importance of beta(2)adrenergic receptor genotype, gender and race on

albuterol-evoked bronchodilation in asthmatics. *Pulm Pharmacol Ther*, 13(3):127–134, 2000.

[83] J. P. Linge, M. A. Williams, C. A. Spronk, A. M. Bonvin, and M. Nilges. Refinement of protein structures in explicit solvent. *Proteins*, 50(3):496–506, Feb 2003.

[84] R. J. Livingston, A. von Niederhausern, A. G. Jegga, D. C. Crawford, C. S. Carlson, M. J. Rieder, S. Gowrisankar, B. J. Aronow, R. B. Weiss, and D. A. Nickerson. Pattern of sequence variation across 213 environmental response genes. *Genome Res*, 14(10A):1821–1831, Oct 2004.

[85] H. Lu, T. Higashikata, A. Inazu, A. Nohara, W. Yu, M. Shimizu, and H. Mabuchi. Association of estrogen receptor-alpha gene polymorphisms with coronary artery disease in patients with familial hypercholesterolemia. *Arterioscler Thromb Vasc Biol*, 22(5):817–823, May 2002.

[86] R. Lüthy, J. U. Bowie, and D. Eisenberg. Assessment of protein models with three-dimensional profiles. *Nature*, 356(6364):83–85, Mar 1992.

[87] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic Acids Res*, 33(Database issue):54–58, Jan 2005.

[88] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic Acids Res*, 35(Database issue):26–31, Jan 2007.

[89] P. Markiewicz, L. G. Kleina, C. Cruz, S. Ehret, and J. H. Miller. Genetic studies of the lac repressor. xiv. analysis of 4000 altered escherichia coli lac repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. *J Mol Biol*, 240(5):421–433, Jul 1994.

[90] A. M. Mathiowetz, A. Jain, N. Karasawa, and W. A. Goddard. Protein simulations using techniques suitable for very large systems: the cell multipole method for nonbond interactions and the newton-euler inverse mass operator method for internal coordinate dynamics. *Proteins*, 20(3):227–247, Nov 1994.

[91] EU 3-D Validation Network. Who checks the checkers? four validation tools applied to eight atomic resolution structures. eu 3-d validation network. *J Mol Biol*, 276(2):417–436, Feb 1998.

[92] F. Melo and E. Feytmans. Assessing protein structures with a non-local atomic interaction energy. *J Mol Biol*, 277(5):1141–1152, Apr 1998.

[93] N. Möckli and D. Auerbach. Quantitative beta-galactosidase assay suitable for high-throughput applications in the yeast two-hybrid system. *Biotechniques*, 36(5):872–876, May 2004.

[94] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–540, Apr 1995.

[95] S. Nabuurs, C. Spronk, E. Krieger, H. Maassen, G. Vriend, and G. Vuister. Quantitative Evaluation of Experimental NMR Restraints. *J. Biomol. NMR*, 3:55–66, 1993.

[96] S. B. Nabuurs, A. J. Nederveen, W. Vranken, J. F. Doreleijers, A. M. Bonvin, G. W. Vuister, G. Vriend, and C. A. Spronk. Dress: a database of refined solution nmr structures. *Proteins*, 55(3):483–486, May 2004.

[97] S. H. Nagaraj, R. B. Gasser, and S. Ranganathan. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief Bioinform*, 8(1):6–21, 2007.

[98] A. J. Nederveen, J. F. Doreleijers, W. Vranken, Z. Miller, C. A. Spronk, S. B. Nabuurs, P. Güntert, M. Livny, J. L. Markley, M. Nilges, E. L. Ulrich, R. Kaptein, and A. M. Bonvin. Recoord: a recalculated coordinate database of 500+ proteins from the pdb using restraints from the biomagresbank. *Proteins*, 59(4):662–672, Jun 2005.

[99] T. Olsson and T. Oprea. Cheminformatics: a tool for decision makers in drug discovery. *Curr Opin Drug Discov Devel*, 4(3):308–13, 2001.

[100] A. Panchenko, A. Marchler-Bauer, and S. H. Bryant. Threading with explicit models for evolutionary conservation of structure and sequence. *Proteins*, Suppl 3:133–140, 1999.

[101] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8):2444–2448, Apr 1988.

[102] Pelletier, Huguette and Kraut, Joseph. Crystal structure of a complex between electron transfer partners, cytochrome c peroxidase and cytochrome c. *Science*, 258(5089):1748–1755, dec 1992.

[103] M. F. Perutz. A tentative atomic model of horse oxyhaemoglobin. *J. Mol. Biol*, 13(2):284–295–, 1965.

[104] J. W. Ponder and D. A. Case. Force fields for protein simulations. *Adv Protein Chem*, 66:27–85, 2003.

[105] A. R. Poteete, D. Rennell, and S. E. Bouvier. Functional significance of conserved amino acid residues. *Proteins*, 13(1):38–40, May 1992.

[106] G. Pugalenthi, K. Shameer, N. Srinivasan, and R. Sowdhamini. HARMONY: a server for the assessment of protein structures. *Nucl. Acids Res.*, 34(suppl2):W231–234, 2006.

[107] G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J Mol Biol*, 7:95–99, Jul 1963.

[108] S. Ranganathan. Bioinformatics education perspectives and challenges. *PLoS Comput. Biol*, 1:e52, 2005.

[109] J. Reboul, P. Vaglio, J. F. Rual, P. Lamesch, M. Martinez, C. M. Armstrong, S. Li, L. Jacotot, N. Bertin, R. Janky, T. Moore, J. R. Hudson, J. L. Hartley, M. A. Brasch, J. Vandenhaute, S. Boulton, G. A. Endress, S. Jenna, E. Chevet, V. Papasotiropoulos, P. P. Tolias, J. Ptacek, M. Snyder, R. Huang, M. R. Chance, H. Lee, L. Doucette-Stamm, D. E. Hill, and M. Vidal. C. elegans orfeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat Genet*, 34(1):35–41, May 2003.

[110] W. Rieping, M. Habeck, and M. Nilges. Inferential structure determination. *Science*, 309(5732):303–306, Jul 2005.

[111] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, Sep 1996.

[112] D. S. Roos. COMPUTATIONAL BIOLOGY: Bioinformatics–Trying to Swim in a Sea of Data. *Science*, 291(5507):1260–1261, 2001.

[113] J.-F. Rual, T. Hirozane-Kishikawa, T. Hao, N. Bertin, S. Li, A. Dricot, N. Li, J. Rosenberg, P. Lamesch, P.-O. Vidalain, T. R. Clingingsmith,

J. L. Hartley, D. Esposito, D. Cheo, T. Moore, B. Simmons, R. Sequerra, S. Bosak, L. Doucette-Stamm, C. Le Peuch, J. Vandenhaute, M. E. Cusick, J. S. Albala, D. E. Hill, and M. Vidal. Human ORFeome Version 1.1: A Platform for Reverse Proteomics. *Genome Res.*, 14(10b):2128–2135, 2004.

[114] J. F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, Oct 2005.

[115] E. Saccenti and A. Rosato. The war of tools: how can nmr spectroscopists detect errors in their structures? *Submitted.*

[116] C. Sachse, J. Brockmöller, S. Bauer, and I. Roots. Cytochrome p450 2d6 variants in a caucasian population: allele frequencies and phenotypic consequences. *Am J Hum Genet*, 60(2):284–295, Feb 1997.

[117] T. Sakaeda, T. Nakamura, and K. Okumura. Pharmacogenetics of mdr1 and its impact on the pharmacokinetics and pharmacodynamics of drugs. *Pharmacogenomics*, 4(4):397–410, Jul 2003.

[118] A. Sali, L. Potterton, F. Yuan, H. van Vlijmen, and M. Karplus. Evaluation of comparative protein modeling by modeller. *Proteins*, 23(3):318–326, Nov 1995.

[119] J. Schultz, R. R. Copley, T. Doerks, C. P. Ponting, and P. Bork. Smart: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res*, 28(1):231–234, Jan 2000.

[120] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res*, 29(1):308–311, Jan 2001.

[121] G. E. Sims and S. H. Kim. A method for evaluating the structural quality of protein models by using higher-order phi-psi pairs scoring. *Proc Natl Acad Sci U S A*, 103(12):4428–4432, Mar 2006.

[122] L. Skrabanek and F. Campagne. Tissueinfo: high-throughput identification of tissue expression profiles and specificity. *Nucleic Acids Res*, 29(21):102–102, Nov 2001.

[123] G. W. Snedecor and W. G. Cochran. *Statistical Methods. Ames.* Iowa State University, USA, 1980.

[124] D. A. Snyder, A. Bhattacharya, Y. J. Huang, and G. T. Montelione. Assessing precision and accuracy of protein structures derived from nmr data. *Proteins*, 59(4):655–661, Jun 2005.

[125] D. A. Snyder and G. T. Montelione. Clustering algorithms for identifying core atom sets and for assessing the precision of protein structure ensembles. *Proteins*, 59(4):673–686, Jun 2005.

[126] C. A. Spronk, S. B. Nabuurs, A. M. Bonvin, E. Krieger, G. W. Vuister, and G. Vriend. The precision of nmr structure ensembles revisited. *J Biomol NMR*, 25(3):225–234, Mar 2003.

[127] E. G. Stein, L. M. Rice, and A. T. Brünger. Torsion-angle molecular dynamics as a new efficient tool for nmr structure calculation. *J Magn Reson*, 124(1):154–164, Jan 1997.

[128] Z. G. Su, S. Z. Zhang, Y. P. Hou, T. Li, D. W. Nebert, L. Zhang, D. J. Huang, L. C. Liao, and C. Y. Xiao. Single-nucleotide polymorphisms in the lipoprotein lipase gene associated with coronary heart disease in chinese. *Eur J Pharmacol*, 454(1):9–18, Nov 2002.

[129] J. Suckow, P. Markiewicz, L. G. Kleina, J. Miller, B. Kisters-Woike, and B. Müller-Hill. Genetic studies of the lac repressor. xv: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. *J Mol Biol*, 261(4):509–523, Aug 1996.

[130] N. Sung, J. Gordon, G. Rose, E. Getzoff, S. Kron, D. Mumford, J. Onuchic, N. Scherer, D. Sumners, and N. Kopell. SCIENCE EDUCATION: Enhanced: Educating Future Scientists, 2003.

[131] S. Sunyaev, V. Ramensky, and P. Bork. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet*, 16(5):198–200, May 2000.

[132] H. L. Tai, M. Y. Fessing, E. J. Bonten, Y. Yanishevsky, A. d'Azzo, E. Y. Krynetski, and W. E. Evans. Enhanced proteasomal degradation of mutant human thiopurine s-methyltransferase (tpmt) in mammalian cells: mechanism for tpmt protein deficiency inherited by tpmt*2, tpmt*3a, tpmt*3b or tpmt*3c. *Pharmacogenetics*, 9(5):641–650, Oct 1999.

[133] J. D. Thompson, D. G. Higgins, and T. J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–4680, Nov 1994.

[134] B. Titz, S. Thomas, S. V. Rajagopala, T. Chiba, T. Ito, and P. Uetz. Transcriptional activators in yeast. *Nucleic Acids Res*, 34(3):955–967, 2006.

[135] S. Tosatto and R. Battistutta. Tap score: torsion angle propensity normalization applied to local protein structure evaluation. *BMC Bioinformatics*, 8(1):155, 2007.

[136] S. C. Tosatto. The victor/frst function for model quality estimation. *J Comput Biol*, 12(10):1316–1327, Dec 2005.

[137] S. J. Triezenberg. Structure and function of transcriptional activation domains. *Curr Opin Genet Dev*, 5(2):190–196, Apr 1995.

[138] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, R. Kent Wenger, H. Yao, and J. L. Markley. Biomagresbank. *Nucleic Acids Res*, Nov 2007.

[139] J. L. Urbauer, M. F. Simeonov, R. J. Urbauer, K. Adelman, J. M. Gilmore, and E. N. Brody. Solution structure and stability of the anti-sigma factor asia: implications for novel functions. *Proc Natl Acad Sci U S A*, 99(4):1831–1835, Feb 2002.

[140] M. Vidal, R. K. Brachmann, A. Fattaey, E. Harlow, and J. D. Boeke. Reverse two-hybrid and one-hybrid systems to detect dissociation of protein-protein and DNA-protein interactions. *Proceedings of the National Academy of Sciences*, 93(19):10315–10320, 1996.

[141] M. Vidal and P. Legrain. Yeast forward and reverse 'n'-hybrid systems. *Nucl. Acids Res.*, 27(4):919–929, 1999.

[142] P. O. Vidalain, M. Boxem, H. Ge, S. Li, and M. Vidal. Increasing specificity in high-throughput yeast two-hybrid experiments. *Methods*, 32(4):363–370, Apr 2004.

[143] A. J. Walhout and M. Vidal. A Genetic Strategy to Eliminate Self-Activator Baits Prior to High-Throughput Yeast Two-Hybrid Screens. *Genome Res.*, 9(11):1128–1134, 1999.

[144] A. J. Walhout and M. Vidal. High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods*, 24(3):297–306, Jul 2001.

[145] M. Walhout, H. Endoh, N. Thierry-Mieg, W. Wong, and M. Vidal. A model of elegance. *Am J Hum Genet*, 63(4):955–961, Oct 1998.

[146] M. Wiederstein and M. J. Sippl. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucl. Acids Res.*, 35(suppl2):W407–410, 2007.

[147] X. Yang, H. Y. Chang, and D. Baltimore. Essential role of ced-4 oligomerization in ced-3 activation and apoptosis. *Science*, 281(5381):1355–1357, Aug 1998.

[148] K. Young and B. Ozenberger. Investigation of ligand binding to members of the cytokine receptor family within a microbial system. *Ann NY Acad Sci*, 766:279–81, 1995.

[149] Z. Zhao and E. Boerwinkle. Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res*, 12(11):1679–1686, Nov 2002.

[150] Z. Zhao, Y. X. Fu, D. Hewett-Emmett, and E. Boerwinkle. Investigating single nucleotide polymorphism (snp) density in the human genome and its implications for molecular evolution. *Gene*, 312:207–213, Jul 2003.

[151] L. Zhu, E. Harlow, and B. D. Dynlacht. p107 uses a p21cip1-related domain to bind cyclin/cdk2 and regulate interactions with e2f. *Genes Dev*, 9(14):1740–1752, Jul 1995.

[152] E. Zuckerkandl and L. Pauling. *Molecular diseases, evolution, and genic heterogeneity*, volume 1 of *Horizons in Biochemistry*. Academic Press, 10 Jan. 1962.

Acknowledgements

First of all I wish to acknowledge Professor Ivano Bertini for trusting and supporting me during these exciting years at CERM. It has been indeed an invaluable school of life and science and I'm sure it will be really fruitful. Discussing about science (and not only) with him is always a surprising experience. I would like to acknowledge Professor Claudio Luchinat for the valuable insights and observations he contributed to me during these years and while writing this thesis, a thanks is also due to Professor Antonio Rosato.

I want to thanks all *cermians*. They are too many to be listed, but I want to acknowledge all of them for big and small suggestions, comments, and contributions to my works.

Thanks to Leonardo Tenori, great friend and science companion: we shared great times, so I think we are going to keep this way! A big thanks also to my friend Barbara and Elisabetta for always encouraging me.

I also wish to acknowledge Julian Lange for expressing the desire of proofreading this thesis before the final submission. He actually did before realize he was going to be late for a squash match; so he kicked my thesis away and left. He's always have been a great kicker.

Once again, last but not least, a special thank to my parents and my brother: without their support none of this could have been possible.

Florence, December 27, 2007

This document has been typesetted using MikTeX distribution (release 2.7)
and TeXnicCenter 1 Beta 7.01.