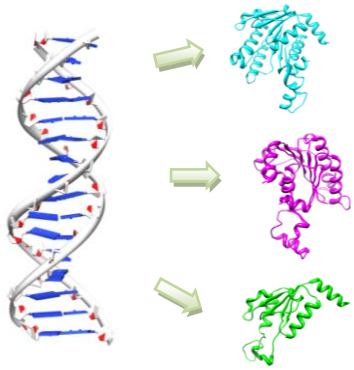
SAPIENZA
UNIVERSITÀ DI ROMA

DOTTORATO DI RICERCA IN BIOCHIMICA
CICLO XXIV (A.A. 2008-2011)

Identification of molecular targets in the human genome



Docente guida
Prof. Anna Tramontano

Coordinatore
Prof. Paolo Sarti

Dottorando
Guido Leoni

Dicembre 2011

To my father

Acknowledgments

There are too many reasons to thank the people mentioned in this page and all the pages of the present work would not be sufficient to list them all.

I would like to thank all the members of the Biocomputing group at Sapienza University. There are many small memories that bind me to each of them.

Moreover I would like to thank my colleagues at INRAN and Dr. Fabio Virgili. Part of this work has greatly benefited from our interaction and the time I spend with them every day is socially and scientifically rewarding.

I am grateful also to Dr. Fabrizio Ferre and Dr. Domenico Raimondo for the many times during my Phd in which they have patiently listened and answered my thousand questions regardless of the fact that it was night or Sunday.

I am particularly grateful and indebted to Dr. Loredana Le Pera. Much of this work would not have been possible without her advices, her patience and her friendship.

Finally my deepest appreciation is for Professor Anna Tramontano. She introduced me to bioinformatics and everything I have learned professionally originates from her teachings. Above all she taught me to have the courage to believe a little more in myself and this is the main thing that I will always remember in my life. It has been an honor for me to have had the opportunity to do my Phd under her supervision and I will do my utmost to demonstrate that her confidence was well placed

1 Introduction

The advent of the post-genomic era and the development of new high-throughput analytical technologies led to firmly establish that several molecular and environmental components, like nutritional molecules, environmental pollutants, chemical compounds with biochemical activity, can have a wide spectrum of biological activities, acting on different products of the genome.

This consideration has important implications in several fields ranging from molecular biology to nutrition and pharmaceutical chemistry; for example leading to rediscover a growing number of pharmacological molecules for being active also towards new targets resulting in new therapeutic effects or sometimes in previously unrecognized side effects[123].

Consequently, the approach for understanding the mechanisms that regulate biological processes was gradually modified for answering to the "omic" need of considering all the components expressed and modulated by a stimulating factor in different diseases or biological conditions.

As a consequence over the past ten years the development of disciplines such as pharmacogenomic was speeded up with the main objective of evaluating the correlation between gene expression and drug response in patients by associating different gene profiles with a drug's efficacy or toxicity.

The development of the “omics” disciplines grew in parallel with the development of the “omic” analytical technologies and it is now possible to completely capture the transcriptome and partially the proteome of a cell upon different treatments[456].

The great power of the methods for quantitatively and qualitatively measure the mRNA and the proteins produced under specific biological conditions is limited by both the intrinsic technical properties of the respective technologies and by the low correlation between the types and the levels of mRNA dynamically expressed by a genome with respect to the

correspondent observed proteins.

This results in a general difficulty in reconstructing the integrated map of the molecular components the expression of which is induced at the mRNA level and that are further translated in functional proteins[789].

The high dimensionality of the produced datasets adds another level of complexity due to the high number of hypothetical target networks modulated by a treatment that can lead to the observed phenotype, resulting in the impossibility of testing all of them with experimental approaches[10.].

It is possible to reduce the dimensionality of the data produced by omic technologies by screening of biological expression networks with multi-target docking studies[11] or predicting protein-protein[12]and protein-ligand[13] interactions by means of structure similarities.

However, also in this case the applicability of methods based on computational screening is limited by the number of crystallographic structures deposited in PDB database (to date: 1699 structures of human proteins) compared to the number of human protein coding genes (21823 source Ensemble rel62).The disproportion between the known structures and the number of genomic products results in the impossibility of covering all the proteins functions and shapes encoded by the genome.

Moreover it is necessary to further consider that 95% of the human genes are subjected to alternative splicing so that a gene can express different mRNA isoforms as a consequence of specific cellular signals.

The transcribed isoforms may have low sequence similarities between them and might not give raise to proteins because of low stability of their mRNA or because targeted by the NMD pathway.

Finally, even if correctly translated they might produce proteins with different ability of interacting with their biological counterparts or ligands[14 15].

A correct assessment of the real capability of the expressed isoforms to encode for functional proteins would improve the sensitivity of the predictions produced by computational screenings of molecular targets.

Unfortunately this task cannot be unambiguously answered by

experimental approaches. While the presence of a functional protein in the cell can be demonstrated, it is impossible to assess that a given peptide sequence is not present or functional at any given time or in any compartment of a cell or an organism.

In this case the question is whether computational strategies, relying on structural and functional predictions, can help to infer the functional roles of the products of transcription.

If properly used, these strategies could play a critical role in investigating protein molecular function and interactions.

Moreover the combination of the assessment of the functional role of the expressed products of transcription with biological informations about the structure and dynamics of the underlying pathways and networks could also permit to better comprehend the molecular mechanisms that regulate biological processes modulated by an environmental or pharmacological stimulus.

Problems associated with the experimental and computational tools to be combined are reviewed in the following paragraph.

1.2 Dynamic nature of the human genome

The human genome is a complex structure organized in transcriptional units clustered in specific regions of the chromosomes (genes) separated by large intergenic non coding regions[16].

Even if the entire nucleotidic sequence of the draft of our genome is known, the functional organization of its transcriptional units is not yet firmly established.

This is mainly due to the dynamic and complex nature of the biological mechanisms that regulate gene expression and to the way in which genes and transcripts are identified on the genome

1.2.1 Gene definition methods for predicting the exon-intron structure of genes

The complexity of the genomic regulatory mechanism makes it impossible to annotate the complete set of transcripts that can be expressed only by means of experimental data.

Current strategies for functional annotation of the genome combine the biological information deriving from cDNA expressed sequence tags (ESTs) or known protein sequences with computational strategies that identify specific nucleic acid sequence motifs that are recognized by the cellular machinery responsible for transcribing, processing and translating messenger RNA molecules.

The minimal set of signals that describe the structure of a coding sequence (CDS) includes the start and stop codons and the donor and the acceptor splice sites for each intron.

Generally these signals are modeled as position specific scoring matrices (PWM) derived from the alignment of functionally related sequences attesting the probability of observing a nucleotide in a specific position of the signal motif.

Moving a window of the same size of the signal and associating the corresponding probability scores, it is possible to predict the portions of the genomic sequence that can originate exons and introns.

1.2.2 Alternative splicing

The determination of the functional organization of a gene is given by the way in which the exons are assembled within the genomic portion to give, once translated, alternative transcripts.

The determination of the total number of alternative transcripts that the alternative splicing machinery could produce is obtained by mean of computational predictions based on *hidden markov model* (HMM) or *support vector machines* (SVM).

Both approaches utilize different features such as splicing signals located within the predicted exons and introns[17][18], phylogenetic informations[19] or include expression data from exon arrays, RNAseq or SAGE expression data[20]

1.2.3 Genomic databases

The effort to centralize the information derived from the annotated sequences and associated annotations produced by the whole genome sequencing projects led to the creation of genome browsers.

The best examples are the three fully established whole-genome browsers: the NCBI map viewer, the UCSC genome browser and the ENSEMBL browser at Sanger Institute.

Each of these present by default a set of contributed gene-finding predictions from different programs and for each new released assemblies.

In addition, each site develops its own gene set based on mRNA evidence obtained from cDNA and EST sequences supplemented by computational predictions.

A fairly conservative database is ENSEMBL in which only genes supported by experimental evidence of at least one isoform via sequence homologyare included.

Apart from these, there exist several genome-browser (ASDb, HDBAS, ASPIC) dedicated to collect information from algorithms specifically trained on alternative splicing genes.

Independently from the considered database, many of the annotated transcripts are originated by computational pipelines therefore with a degree of uncertainty in their determination due to the accuracy limits ofthe used prediction algorithms.

Periodically, with the improvement of splicing prediction algorithms or after the release of new versions of the genome, the functional organization of entire genes radically changes leading to the disappearing or to the

rearrangement of the genomic sequence of their isoforms.

1.2.4 Factors regulating the expression of the coding RNA

Expression of transcription products is regulated at the transcriptional and post transcriptional level.

The transcriptional regulationis realized by transcription factors (TFs), such as the TBP-associated factors, or TAF proteins that dynamically act on a wide variety of cis-acting regulatory elements[21 22 23 24]

Post-transcriptional regulation alters the levels of mRNA expressed by the transcriptional regulatory mechanisms and is carried out by several systems.

In first instance, specific regulatory small RNAs can modulate gene expression by interacting with target mRNAs favoring their degradation or modulating their translation.

Second, the stability[25] of the tridimensional structure of the mRNA itself influences the decay rate of expressed transcripts, reducing the affinity of unstable mRNAs for proteins deputed to regulate the translation of RNA, and post-transcriptional modifications, such as RNA splicing, and editing[26 27]

Finally, enzymatic control mechanisms limit the possibility to encode for non functional proteins binding to functional motifs in the untranscribed regulatory region or monitoring the presence of premature stop codons in the coding region, resulting in the consequent degradation of faulty mRNA by biological pathways such as the non sense mediated decay pathway.

These mechanisms are not able to remove all the non functional mRNAs and, in several cases, a reduced amount of transcript is translated in a non functional protein[28] with important implications in the development of diseases.

Several computational methods have been developed to infer mRNA stability or its propensity to be targeted by non sense mediated decay and recent estimates report that the proportion of potential non sense

alternatively spliced transcripts in human genome is roughly 10%[29]

Therefore a considerable fraction of isoforms (90%) remains that is reasonably not directly targeted by degradation pathways but still modulated in its dynamic expression.

Expression profiling experiments usually take a snapshot of the mRNA levels in the cell and do not capture the dynamics of mRNA synthesis and breakdown.

This lack of information on the dynamic component of the regulation of mRNA levels limits biological investigations in general and proper modeling of transcriptional networks in particular.

1.3 Computational approaches for building omic network and screening of molecular target

The biology of an organism (cell, animal, plant) results from the joint operation of a large network of biochemical reactions, catalyzed by the collaborative action of enzymes encoded in the genome of that organism.

These reactions form routes that can be grouped in pathways surveying different biological functions.

The effect of biological, chemical or pharmacological compounds is carried out through the capability to indirectly or directly modulate proteins belonging to biological pathways

It is therefore possible to extrapolate the phenotypic effect of a treatment from the genotypic expression by evaluating the number of biological pathways modulated in an omic set of expression data[30].

Even if several databases and tools[31] have been developed to assist biologists in the extraction and building of the modulated interaction networks underlying biological pathways (some of them were utilized in the paper III and IV), the power of network reconstruction approach is limited by the consideration that the number of experimentally determined interactions stored in public databases is an approximations of the real number of

interactions taking place among the products of the genome.

This results in the reconstruction of incomplete networks with large groups of not connected proteins/genes making it hard to reconduct the modulated genotype to the observed phenotype.

To enrich the network built by data mining of experimental interactions annotated in biological databases, it is possible to use computational approaches for predicting new likely interactions not yet characterized at the experimental level.

Some approaches detect interacting features or domains evolutionarily conserved in pairs of proteins with unknown interactions[32].

Others simulate the protein-protein[33] or the protein-ligand[34] binding (docking) by mean of estimates of the differential free energy between the bound and unbound conformation for each possible complex, considering favorable the complexes with negative differences of their free energy.

The key requisite, which is also the limit of the application of docking analyses for multi target screening studies, is the availability of protein structures to be utilized in docking simulations.

As of August 2011 the PDB database contains approximately 1700 human structures and only 14 human proteins have structures for at least two of their alternatively spliced forms[35].

The reduced number of PDB structures with respect to the number of genes or proteins that can be a potential target in a network originated from expression profile experiments is still inadequate to explore the search space of all the possible ligand protein interactions.

1.4 Methods for protein structure and function prediction

The functionality of a protein is strictly related to its structure.

With the help of modern computational methods it is possible to predict the structure of a protein and to also evaluate the presence of features related to specific biological functions.

The applicability of structural and function prediction methods is limited in part by the lack of tools that deal with the problem atthe genome scale and by the difficulty to evaluate the efficacyof the obtained predictions in discriminating between functional and non functional isoforms within a gene.

1.4.1 Protein structure prediction

Basically, structural prediction methods are divided in ab initio methods that build models on the basis of physicochemical rules and search for the most thermodynamically stable structure under physiological conditions,and methods that build structures (called targets) on the basis of their homology with evolutionary related proteins thestructure of whichhas been experimentally determined (named template).

The application of both approaches to annotate the genome has several bottlenecks.

The first approach is particularly expensive in terms of computational resources making extremely difficult to apply these methods to model entire genomes and is as yet feasible only for very small proteins.

The homology modeling methods are less expensive in terms of computational resources and potentially easy to apply atthe genomic scale but their application is limited primarily by the availability of deposited crystallographic structures (templates).

Moreover, it is necessary to consider that the reliability of the homology models is strongly influenced by the accuracy in the alignment between target and template sequences in order to correctly detect structurally and evolutionary related residue pairs.

Often, the existing automated pipelines for building models allow user provided sequence alignments to be used in order to give the possibility to the user to manually check and improve the overall quality of the alignment and/or the selection or removal of templates.

This manual approach is useful when a limited number of models are required but unfeasible when the number of models to be obtained is at the genomic scale.

1.4.2 Functional annotation of proteins

The rapidly increasing amount of proteins sequence data make it impossible to functionally annotate all the produced sequences only on the basis of experimental evidences, therefore most functional annotations are produced by computational methods that recognize features at the primary structure level.

Functional annotations tools can be divided in tools that detect cellular localization signals or post-translational modifications and tools that predict biological functions according to the alignment of protein sequences with sequence profiles of modular protein domains[36] or short functional motifs[3738].

Profiles are elaborated by classifying protein sequences accordingto thermodynamic[39], structural[40], evolutionary[41] and functional factors[42].

The tools that evaluate the presence of functional domains assign the presence of domains on the basis of the local alignment of the aminoacid sequence with profiles of sequences with known domains.

Splicing isoforms often differ for limited portion of their sequence due to the insertion or removal of limited number of exons, therefore commonly used tools can assign functional domains even if they are represented by small portions of sequences, i.e. incomplete. This aspect limits the sensitivity of the functional annotation pipelines posing the question of which is the minimum portion of a functional domain necessary for considering the domain structured and functional.

1.5 Computational analysis of ENCODE data

In 2007 the pilot phase of the encode project,aimed at identify all the functional elements in 44 selected regions that make up 1% of the human genome,was completed.

The obtained results[43] highlighted several important features of the human genome, remarking the central role of alternative splicing in regulating gene expression and detailing a reference set of manually annotated splice variants by the GENCODE consortium[44].

The in depth computational analysis of the GENCODE transcripts was the first attempt to evaluate the coding potential of alternative splicing isoforms and suggested that a large number of the analyzed splice variants (around 50%) are likely to encode for proteins with potentially deleterious changes in their protein structure and function.

In the following years other studies addressed the same question with controversial conclusions.

Some evidences reported splicing events that produce variants showing novel and sometimesunexpected structural and functional properties[45]compared to their native counterparts. Others reported that splicing events not conserved across different species are more likely to produce protein with big rearrangement of their sequences resulting in unstable conformations[46].

These findings raised several interesting questions, but also a few practical issues.First of all, the careful manual analysis performed on not more than the 1% of the genome needs to be scaled up to the whole genome and therefore automated. Secondly, there’s a global need of user-friendly computational tools that can assist biologists to setup and analyze their omic experiments.

1.6 Aim and contribution of the study

The purpose of the study described here is to contribute to the field of functional genomics by developing, testing and applying computational methods to the problem of the evaluation of the effects of environmental and pharmacological molecules on genome expression.

Original results are described in four independent papers that address the general problem of identifying potential pharmaceutical targets and study their susceptibility to external interfering agents.

The first two papers are devoted to the investigation of the coding potential of alternative splicing products in the human genome. As previously mentioned, the assessment of the real capability of the expressed isoforms to code for functional proteins is of fundamental importance because would permit to gain insights in the biological meaning of the expression data produced by the omic technologies and to improve the sensitivity of the predictions produced by computational screenings of molecular targets.

In the first paper, taking advantage of omic databases, we assessed the power of computational strategies to infer the functional roles of the products of transcription.

The second paper describes the implementation and benchmark of the MAISTAS structural prediction server which is able to analyze the structural plausibility of the isoforms produced by alternative splicingat a genomic scale.

Papers three and four are devoted to the application of computational techniques to investigate the molecular targets and the effects on their expression of molecules known to interfere with the physiological functions of a cell.

In particular these techniques were applied on a class of compounds (tocotrienols) constituents of the Vitamin E.

Paper three explores the possibility that specific classes of tocotrienols bind specific subtypes of the nuclear estrogen receptor taking

advantage of docking studies followed by experimental validation.

Paper four investigates the molecular mechanism leading to the arrest of cell cycle in breast cancer cells after treatment with specific tocotrienols by analyzing microarray experiments and the implied pathways and networks.

This second part of the work was carried out at I.N.R.A.N. in the group of Dr Fabio Virgili and the biological experiments analyzed and mentioned in the present work were produced by Dr.ssa Raffaella Comitato and Dr. Roberto Ambra.

2 Results

This chapter shortly outlines the main results obtained during this work and detailed in the research papers.Results are divided in two sections.

The first section is devoted to the general problem of the evaluation of the coding potential of alternative splicing. The second section reports the application of computational methods to evaluate the capability of bioactive molecules constitutive components of vitamin E, to modulate the human genome expression.

Findings in each area are presented in the context of related research.

2.1 Capability of computational methods to infer the functional role of the products of transcription

Alternative splicing is the preferred mechanism used to quantitatively control the gene expression and functionally diversify proteins produced by the same gene.

Estimates of the amount of alternative splicing in human have risen dramatically over recent years, especially since the advent of high throughput transcriptomic sequencing[474849] technologies, reaching up the 95% of the multi-exons genes[50].

Despite the increasing of the overall number of sequences annotated in genomic databases is still unclear the exact functional role of protein isoforms produced by splicing.

Modern proteomic analysis techniques allow the identification, characterization and quantization of up to several thousands of proteins in a single experiment[51].

Therefore it is theorically possible to compare transcriptomic and proteomic expression profiles in order to assess the correlation between

mRNA and protein levels in the system of interest.

Although the detection of a proteomic peptide identifying an isoform does not conclusively ensures that it is functional, it does imply that the corresponding transcript is translated into a protein likely to fold and be produced at sufficient levels to be detected.

Small scale applications of this concept produced interesting results on limited datasets of 16 human genes[52] and 130 fruit fly genes[53].

Unfortunately, a large scale survey is still difficult to realize because the high throughput proteomic techniques, even if approaching the genomic scale, are still not able to detect and quantify all the peptides produced by protease digest and this hinders the analysis.

Given these premises, computational methods are the only way for obtaining a probabilistic estimate of the likelihood that a protein is functional.

In paper I we assessed how these strategies can help to correctly recognize functional transcription products likely to be translated.

Such estimate has been performed in two steps. First we evaluated how well methods for the prediction of the structural and functional plausibility are able to identify alternative splicing isoforms unambiguously identified by proteomic expression data.

Next we applied the same methods to isoforms of the same genes that are not detected in any publicly available proteomic experiment database, even if potentially detectable.

To build the datasets the coding portion of the alternatively spliced genes stored in Ensemble (ver57) was characterized defining the specific and the aspecific portions of the exonic sequences of each isoform

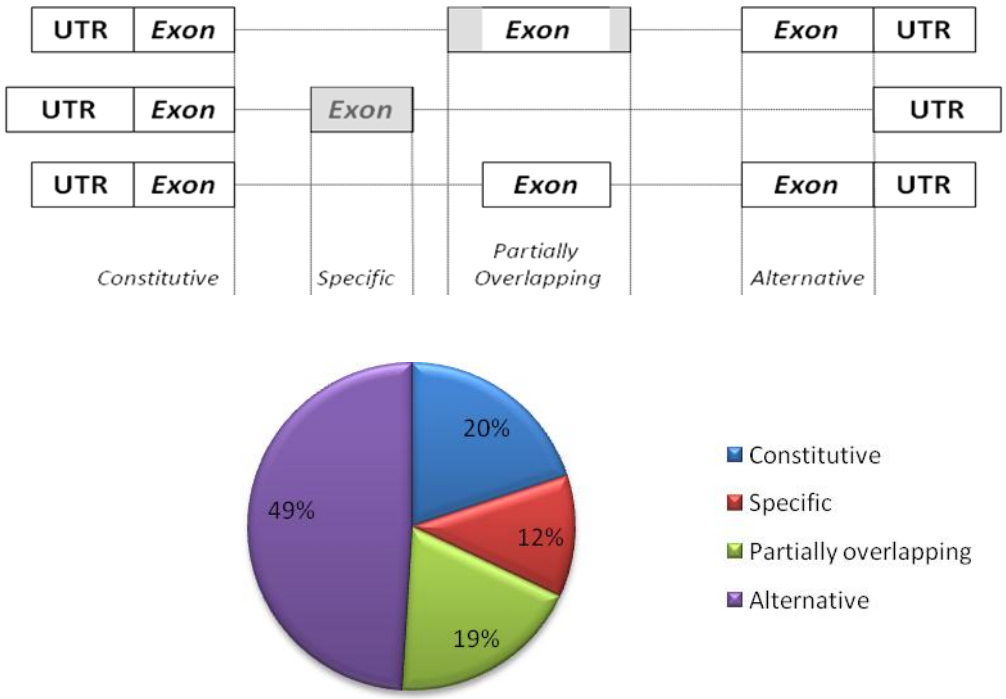


Figure 1: Characterization of the coding portion of human genome (Ensemble57) In the pie chart the percentages of different types of exons in the coding portion of alternatively spliced genes are represented. The specific parts of exonic sequences in isoforms belonging to the same gene are depicted in grey.

Proteomic peptides were mapped on the specific exons included in transcript sequences. Isoforms unambiguously identified by peptides (mappingtotheir specific exons) were included in the positive dataset only if it was possible to select another isoform of the same genenot identified by proteomic peptides but potentially identifiable (i.e. with specific parts in its sequence potentially detectable by mass spectrometry).The detailed pipeline is described in paper I.

The analysis was carried outby evaluating, for each isoform, its structural plausibility, based on structural models built by homology, the

presence of complete domains, based on the Pfam domain definition, the presence of functional sites such as catalytic sites or ligand binding sites, based on SwissProt annotated features[54].

The application of the previously mentioned strategies necessarily implied the establishment of generally valid criteria to assess the likelihood that an isoform is structurally and functionally plausible.

For functional features, such as catalytic sites and ligand binding sites, the criterion was easily established and consisted in verifying the removal of functional sites experimentally determined and annotated in SwissProt database.

The assessment of functional plausibility in terms of the presence of functional domains was more complex.

It is known that alternative Splicing can remove whole protein domains, but tends not to occur within domains[55]. It is therefore reasonable to assume that isoforms with truncated domains are the result of an incorrect event of splicing more than of a regulated event or that they have a completely different function from the cognate isoform with a complete domain.

The first question is which is the minimum portion of a functional domain necessary for considering the domain itself as structured and functional.

To answer to this question, all the human protein sequences for which the structure is known were analyzed; assessing to which extent their sequences cover profiles representing known classes of functional domains.

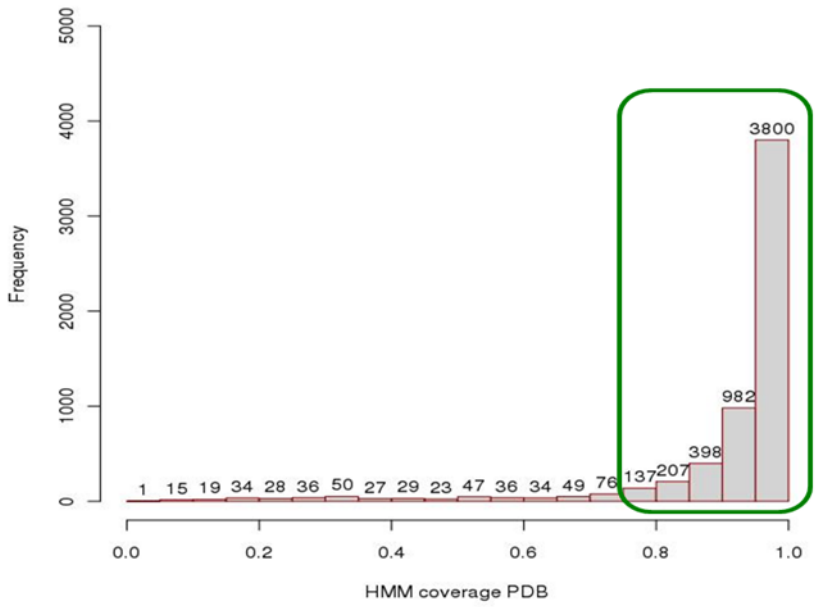


Figure2: Coverage of PFAM domains in PDB structures: The plot shows the percentage of coverage of PFAM domain sequence profiles in 3859 PDB structures of human proteins with less of 30% of sequence homology. The 90% of domains (in green box) cover the corresponding PFAM profile for more than 70%.Therefore this threshold has been chosen for considering a domainas structured and functional

Structural predictions were evaluated by estimating several parameters: the quality of the packing of the core model structure, the extent of the hydrophobic surface exposed to solvent and the presence of insertion and deletions that could not be accommodated without larger modification with respect to the structural template.

The assessment of isoforms plausibility according to the previously described criteria, reveals that the isoforms whose existence is confirmed by proteomic are strongly enriched in “positive” parameters (i.e. parameters that are similar to those observed for experimentally confirmed proteins) than the isoforms not confirmed by proteomics data.

90% of the positive modeled isoforms appear to be structurally plausible or with integer functional domains, while only 40% of the isoforms

in the datasets not confirmed by proteomic have plausible structural and functional parameters.

The combination of structural and functional parameters in each isoform has been used for assessing the sensitivity of the computational strategies in the identification of isoforms produced bya regulated alternative splicing event (summary of the statistical analysis is reported in Table1).

	Coverage	Accuracy	Sensitivity	Specificity
AS	0.14	0.69	0.99	0.39
St	0.26	0.72	0.91	0.52
Pfam	0.81	0.72	0.95	0.41
AS U St	0.35	0.71	0.93	0.50
St U Pfam	0.89	0.71	0.93	0.44
AS U Pfam	0.85	0.71	0.95	0.43
AS U St U Pfam	0.92	0.70	0.93	0.45
AS ∩ St	0.06	0.71	1.00	0.25
St ∩ Pfam	0.18	0.80	0.99	0.49
AS ∩ Pfam	0.10	0.77	1.00	0.28
AS ∩ St ∩ Pfam	0.05	0.80	1.00	0.23

Table 1: Assessment of the sensitivity of structural and functional parameters

The combination of the three criteria described above provides the most accurate prediction being able to identifythe 80% of these isoforms even if this combination is applicable only in a limited fraction of cases (5%).

The most limiting criterion is the detection of the presence of functional sites due to the limited amount of cases for which this feature is annotated.

Considering only if an isoform has not interrupted functional domains or has a plausible structure results in the more useful criterion of evaluation, applicable in most cases with high accuracy.

The absence of a well defined negative set does not allow the estimate of the exact number of false positives i.e. the number of isoforms not translated in proteins but predicted as plausible; however the estimates

obtained on positive isoforms highlight the power of computational methods for inferring the functionality of transcription products.

2.2 Large scale automatic structural evaluation of genome products

The work described in paper I highlights the importance of the structural predictions for inferring isoform functionality.

However, there are still few automatic publicly available instruments that allow structural predictions to be obtained at the genomic scale and most of them only collect models already existing[56] or provide information about the putative effect of alternative splicing provided that at least one of the isoforms has an experimentally solved structure[57].

During this PhD project a publicly available server named MAISTAS (described in Paper II) has been implemented and benchmarked. MAISTAS collects and builds, whenever possible, comparative homology models for all the putative spliced isoforms of the genes in genomes stored in the Ensemble database, providing an estimate of the likelihood that the isoforms correspond to potentially stable and structurally plausible proteins in the absence of major conformational rearrangements.

The assessment of the structural plausibility of the obtained models is performed comparing their structural properties with those observed in known protein structures and in the closest homologs of known structure

More specifically, models are evaluated by assessing the putative effect of deletions and insertions of parts of their sequence compared with the sequence of the closest template, the packing of the core of themodeled structures andthe extent of their exposed hydrophobic surface.

The server was benchmarked on the entire dataset of human alternatively spliced isoform whose existence at the protein level could be unambiguously verified by mass spectrometry according to PeptideAtlas database.

In benchmark tests, MAISTAS was able to produce and analyze models in 30% of the cases. In the 70% of the to be modeled sequences the model could not be built because there is no protein of known structure satisfying the parameterschosen for considering it as a good template.

Out of the modeled isoforms, the 80% was assessed as structurally plausible.In the majority of the remaining cases, the obtained model showed a large hydrophobic surface exposed to the solvent.

In these cases, the protein might indeed represent an incomplete and therefore not plausible structure, but it could also be a subunit of a larger complex.

2.3 Computational methods for the identification of molecular targets: Vitamin E case study

Vitamin E is a generic term used to refer to a family of fat-soluble compounds composed by α, β, γ, δ-tocopherols and corresponding four tocotrienols[58.]

Several studies report that α, γ, δ tocotrienols have a wide spectrum of specific functions and activities (ranging from antioxidant activity to the protection against the stroke and diabetes typeII) distinct from those attributed to tocopherols[59 60]

Moreover in recent years there is evidenceof a role associated with a possible “anti-cancer” activity exhibited by tocotrienols achieved by modulating the cell cycle and inducing apoptosis in prostate and breast cancer cell lines[6162] .

Despite the growing interest about the potential applications of these alimentary compounds as dietary supplement in order to coadjuvate the pharmacological therapies for cancer treatment[63], the mechanism of action underlying their activities and their molecular target(s) are still poorly understood.

The broad spectrum of tocotrienols activities suggests that the molecular target through which their activities are carried out is in a central position in different biological pathways or alternatively that tocotrienols could be able to bind more biological receptors related to different biological functions.

Tocotrienols, therefore, represent a suitable case study in which the computational methods can be used to drive and interpretate experimental results in order to elucidate the mechanism of action of these molecules.

The analysis of expression data from cDNA array experiments previously performed at I.N.R.A.N. on cultured human breast cancer cells[64], in sub-cutaneously implanted athymic mices[65] and on stripped culture mediums (characterized by total removal of lipidcomponents of serum, including estrogens), indicated the possibility that the biological activities of tocotrienols could result from interaction with the estrogen receptor related signaling.

In order to confirm or discard this novel hypothesis docking studies to estimate the theorical propensity of α, γ, δ tocotrienols to bind the two isotypes of the estrogens receptor (ERα, ERβ) in their active and inactive conformation were performed.

The results (detailed in paper III) suggest thattocotrienolscan preferentially bind the ERβ in its active conformation with a binding mode similar tothat of estradiol (the estrogenic hormone, which is the natural ligand of this receptor).

With the help of computational techniques it was also possible to estimate the levels of activity (expressed in terms of the theoric binding energy of the docked compound within the receptor) shared by different tocotrienols, highlighting the decrease of their activity inversely related to the number of methyl groups in their phenolic ring.

The quantitative differences in tocotrienols activity and the preference for a specific receptorial isotype were subsequently confirmed by in vitro binding-displacement test analyses (Figure 3).

Displacement tests cannot provide any information about the ability of these compounds to place the receptor in an active or an inactive conformation. Therefore experiments of indirect immunofluorescence, monitoring the cellular localization of estrogen receptor upon tocotrienols treatment in breast cancer cells (MDA) expressing only the isotype beta at protein level, were performed.

Also in this case the results obtained by docking were confirmed revealing that, upon tocotrienol binding, ERβ is able to translocate into the nucleus of the cell to exert its transcription factor activity similarly to the receptor activated by estradiol.

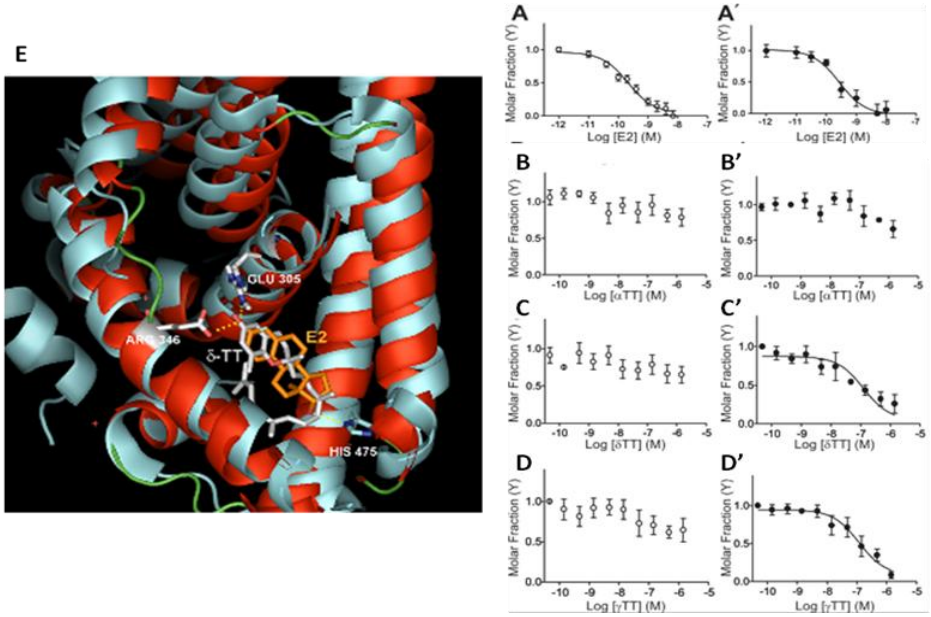


Fig 3:Binding displacement test and docking results: Displacement binding analyses of tested compounds (estradiol, α, γ and δ tocotrienols) versus ERα(A,B,C,D) and ERβ(A',B',C',D'). According to the obtained results only γ and δ tocotrienol were able to bind ERβ. On the left (E) the superposition between crystallographic structures of ERβ (PDB: 1X7J) bound to estradiol (structure in cyan and ligand in orange) and the best dockingcomplex in δ tocotrienol simulationsare shown.

The results of paper III undoubtedly highlighted the potentialities of docking methods for screening molecular targets but also pointed out one limit of the technique: the difficulty to compare and to rank different binding energies in different protein-ligand complexes.

The problem was addressed in the tocotrienol case study using a qualitative evaluation of the solutions produced in each docking. We compared the binding energy of the most favored solutions (by mean of clustering) with those of the reconstructed natural ligand-receptors complexes.

This implies that there is still room for improvement and more effort should be put indeveloping and testing “normalization” procedures of the binding scores across different resultsobtainedby docking experiments.

2.4 Pathway and network analyses: Involvement of the estrogen receptor beta signal transduction in tocotrienol activity

The application of pathway and network analyses to omic expression data can help reconstructing the regulatory interplay related tothe biological activity of a target.

The potential of these methods was explored by studying the mechanism leading to the arrest of cell cycle and apoptosis in breast cancer cells after tocotrienols treatment.

The results obtained and described in paper III highlighted the role of the ERβ as a potential biological target of tocotrienols.

Several studies on cell and animal models report that this receptor is the gateway of several biological pathways and its biological action is the result of an intricate regulatory interplay not completely understood and involving ERα.

The two receptorial isotypes are expressed by different genes and perform partially overlapping biological functions[66].

In several tissues they are co-expressed and usually ERβ,when activated, exhibits an inhibitory action on ERα-mediated gene expression[67].

Besides this “antagonist” relationship between the two receptors, it has also been recently observed that there is a “synergistic” relationship in some biological districts directed to positively modulate the same biological functions.

An example of this is the neuroendocrine tissue in whichERβleads to an increase of the amount of produced oxytocin peptide, while ERα increases expression of its receptor[68].

It follows that the biological response after the activation of ERβ could vary depending on the biological district or condition considered and on the co-expression the two receptors.

Tocharacterize the regulatory gene expression profile induced by different tocotrienols, microarray experiments were performed on breast cancer cells (MCF7) expressing constitutively both ERα and ERβ.

Lists of differentially modulated genes after γ and α tocotrienol treatment versus control were produced and subsequently submitted to network and overrepresented pathway analysis, providing insightson the tocotrienols mechanism of action in MCF7 cells.

Microarrays analysis suggests that, in breast cancer cells, the mechanism of action ofERβ upon tocotrienol treatment is mainly explicated via an antagonistic relationship with Erα-mediated gene expression.

This is highlighted by the observed down regulation of Era expression together with the up regulation of genes encoding for repressors of transcription factors (NRIP1, THR, HEXIM1, MECP2).

According to the pathway and network analysis (detailed in paper IV) these genes belong to a common sub network of interaction involving ERβ.

We also found that, the p53 tumor suppressor (downstream ofErα) gene expression is up regulated.

Evidences report that ERα is able to repress the expression of this cellular death inducing gene and that the lowering of the control on p53expression could increase the signal of cell death.

The decrease in ERα expression and the activation with consequent nuclear translocation of ERβ were also confirmed by immunofluorescence experiments.

Pathway analysis suggested the specific capability of γ tocotrienol to modulate pro-apoptotic genes (CASP10, BID, AIFM1, APAF1). Several mitochondrial genes appeared to be negatively modulated after γ tocotrienol treatment (COX4L1, BCL2A1) together with the SERCA3 ATPase responsible for the reabsorption of cytosolic calcium inside the endoplasmatic reticulum.

Interestingly, the increase of cytosolic calcium is one the major mitochondrial pro-apoptotic signals[69].

All these evidences suggest that the molecular mechanism underlying γ tocotrienol inducedapoptosis in MCF-7 cells is, at least in part, mitochondriadriven.

With the help of omic network analyses it was therefore possible to partially interpretate the gene expression profiles produced by microarrays and assess the presence of genes that belong to pathways likely to be modulated by tocotrienols treatment.

It should also be mentioned, though, that the analysis of the interaction network between modulated genes resulted in many cases difficult to interpretate.

Many modulated genes have context dependent functions and it is not possible to exclude that the change in their expression levels is not directly connected to the treatment.

An example in the presented case study is the observation, after treatment with γ tocotrienol, of the up regulation of few genes related to the development of mammary carcinoma (detailed in paper IV).

Their up regulation could be due an adaptative response to the pro-apoptotic stimulus or to interactions not yet annotated in public databases or be a false positive result of microarrays analysis

Furthermore we cannot exclude that some of the over expressed transcripts exert their biological activity throughone of their alternatively spliced isoform.

We think that a more accurate analysis at the isoform level, now possible thanks to the development of next generation sequencing techniques, together with the application of computational methods for evaluating the plausibility of the expressed isoforms could help to achieve more sound results.

On the other hand the determination of the tridimensional structure of single isoforms could also increase the accuracy of computational methods for investigating protein-ligand or protein-protein interactions allowing the prediction of interactions still not annotated or the identification ofoff-targets responsible for activating biological pathways not directly connected to the main effect observed after tocotrienol treatment.

3 Conclusions and outlook

In order to identify hypothetical molecular targets in genome expression products, it is fundamental to evaluate the functional role of the transcriptional units in which the genome is organized and to consider that bioactive molecules can bind to different targets resulting in different biological activities or side effects in case of molecules with therapeutic activities.

The computational analysis of data produced by the pilot phase of ENCODE project suggested for the first time the great potential of computational methods to investigate the properties of the functional elements produced by the 1% of the genome raising the question if these methods could be used on a genomic scale to elucidate the function of the encoded transcripts and the molecular mechanisms underlying the biological effects subsequent to their modulation.

The present work started from this question with the aim of scaling up the analysis performed on the 1% of the genome to comprehend the functional role of genome products and to highlight the potential and some bottlenecks of currently available tools for the prediction of protein structures and functions at the genomic scale as well as of tools for analyzing ligand-protein and protein-protein interactions in “omic”expression networks.

The application of structural and functional prediction methods to all the genomic products requires the development of tools that permit to model and compare with reasonably and widely applicable criteria a large number of sequences.

The development and the benchmark of the first method to analyze the structural properties of all the alternative splicing isoforms produced by the genome confirms that this is a viable route to start addressing the challenges of the post-genomic era.

The application of the homology modeling for structural prediction is partially limited by the availability of suitable templates. However, it is reasonable to assume that this limitation will be overcome by the development of high-throughput protein crystallization platforms[70] and/or by improvement of techniques that combine ab-initio structure prediction methods with homology modeling of fragments of the query sequence on the basis of local similarities with templates.

This latter approach is a good compromise between the pure ab initio methods that require large computational resources and homology modeling methods that are easier to implement but have a limited coverage.

Some recently developed tools that use such as approach[71] won the last 3 editions of CASP, thecommunity-wide experiment for testing the state-of-the-art of protein structure predictions.

By mean of the characterization of the coding portion of human genome, we have shown that the prediction of the structure and function of alternatively spliced proteins can be instrumental in identifying cases in which transcribed isoforms without experimental evidence of existence as proteins have properties significantly different from isoforms of the same gene that have beenexperimentally confirmed at the protein level.

The high percentage of “unusual” isoforms observed in our study (around 40%) is quite similar to the percentage of isoforms with drastic alterations in their structure and function observed in the computational analysis of ENCODE data (around the 55%) and poses the key question of which could be their functional role.

A first hypothesis could be that the different structural and functional parameters produce a completely different structure and function with respect their spliced counterparts expanding the coding potential of their originating gene. Alternatively the change of their structural and functional properties could be used to modulate the biological activity of the gene products. In these two cases, alternative transcripts would increase the

repertoire of protein functions.

There are few experimental evidences[72] [73]documented in literature supporting this hypothesis, however we think that the large number of observed isoforms with unusual structural and functional parameters makes unlikely that this explanation holds for all of them.

On the other hand, some of these isoforms could encode for functional RNAs[74] or exert a regulatory function interacting and perhaps inhibiting other proteins[75].A fraction of the transcripts could also be originated by “non correct” alternative splicing events and one could speculate that the cell tolerates these “non correct” variants to some extent because produced in low numbers since they could represent *de facto* evolutive reservoirs on which the selection pressure against exon loss or substitution is reduced making large evolutionary changes possible.

It is also possible, and perhaps likely that some of the splicing isoforms produced at high levels are the result of “aberrant” events perhaps linked to diseases[76].

Obviously the phenomenon that we observed admits too many possible explanations and it will not be easy to test all of them (especially since each of them can apply to a subset of the cases) and to provide a conclusive and general answer. Certainly the integration of new large scale experimental data produced by more powerful high throughput technologies together with other computational analyses not only limited to the assessment of functional and structural properties will help interpreting their functional role.

In this perspective the work described here is to be intended as a first step and should be complemented in the near future with the analysis of other functional features that can be modified by alternative splicing, such as the modification of cellular localization signal, protein disorder, propensity to be targeted by decay mechanisms, mRNA stability.

It is clearly important to continue in this direction because the

understanding of the coding potential of eukaryote genomes is certainly required for understanding the molecular mechanisms that regulate biological processes and their modulation by bioactive molecules. This aspect is being actively addressed since it has important implications for human health and nutrition as demonstrated by our study of the molecular targets of bioactive molecules (tocotrienols) constituents of Vitamin E in different breast cancer cellular models and of the effect of the molecules on gene expression. In this case, we used docking techniques combined with an assessment of overrepresented pathways and on the reconstruction of interaction networks from transcriptomic expression profiles.

The rationale of our choice of tocotrienols is due to the observation that, despite the increasing concern about the potential applications of these alimentary compounds as dietary supplement to coadjuvate the pharmacological therapies for cancer treatment, the mechanism of action underlying their activities and their molecular target(s) are still poorly understood.

Interestingly, our computational results highlighted that the products of differential gene expression observed in breast cancer cells after treatment with specific subtypes of tocotrienols are linked to induction of apoptosis.

The analyses of the involved networks and pathways also indicated that several transcription factors, related to induction of apoptosis, modulate part of a network of interaction that includes the estrogen receptor beta.

Docking simulations, confirmed by displacement assays in vitro, indicated that Erβ is a molecular target of tocotrienols that act by binding in its active site.

Since immunofluorescence experiments evidenced the capability of ERβ to translocate in cellular nucleus after tocotrienol treatment, it is conceivable that tocotrienols are able to activate this receptor stimulating its transcriptional activity.

The application of computational methods for the detection of targets of vitamin E underlines the potential but also some of the limitations of the available computational methods. For example, the reconstruction of the network of interactions and the analysis of modulated pathways explains only a fraction of the total number of modulated genes, making it hard to reconduct the overall genotype to the observed phenotype.

It is reasonable to suppose that the more accurate analysis at the isoform level, now possible thanks to the development of next generation sequencing techniques, together with the application of methods for the determination of the tridimensional structure of single isoforms, could increase the accuracy of computational methods for investigating protein-ligand or protein-protein interactions and allow the prediction of as yet unknown interactions and/or the identification of off-targets responsible for activating biological pathways not directly connected to the main effect observed after treatment.

The ambition of understanding of the complexity of the functional content of our genome and of the intricate relationships between the different pathways and networks is probably no within easy reach as of today, but efforts such as those described here can hopefully be useful to reach this important and exciting goal.

4 Pubblications

Paper I: Coding potential of the products of alternative splicing in humans

Leoni et al. Genome Biology 2011, 12:R9
http://genomebiology.com/content/12/1/R9

Genome Biology

RESEARCH Open Access

Coding potential of the products of alternative splicing in human

Guido Leoni[1,2], Loredana Le Pera[1], Fabrizio Ferrè[1], Domenico Raimondo[1], Anna Tramontano[1,3*]

Abstract

Background: Analysis of the human genome has revealed that as much as an order of magnitude more of the genomic sequence is transcribed than accounted for by the predicted and characterized genes. A number of these transcripts are alternatively spliced forms of known protein coding genes; however, it is becoming clear that many of them do not necessarily correspond to a functional protein.

Results: In this study we analyze alternative splicing isoforms of human gene products that are unambiguously identified by mass spectrometry and compare their properties with those of isoforms of the same genes for which no peptide was found in publicly available mass spectrometry datasets. We analyze them in detail for the presence of uninterrupted functional domains, active sites as well as the plausibility of their predicted structure. We report how well each of these strategies and their combination can correctly identify translated isoforms and derive a lower limit for their specificity, that is, their ability to correctly identify non-translated products.

Conclusions: The most effective strategy for correctly identifying translated products relies on the conservation of active sites, but it can only be applied to a small fraction of isoforms, while a reasonably high coverage, sensitivity and specificity can be achieved by analyzing the presence of non-truncated functional domains. Combining the latter with an assessment of the plausibility of the modeled structure of the isoform increases both coverage and specificity with a moderate cost in terms of sensitivity.

Background

Alternative splicing (AS) is a mechanism used by cells to diversify the proteins produced by a gene. Estimates of the amount of AS in human have risen dramatically over recent years, especially since the advent of novel high-throughput sequencing technologies [1-3], reaching up to the 95% of the multi-exon genes [4].

While the role of AS in expanding the functional complexity of a genome is established, less clear is whether all generated transcripts do indeed encode functional proteins and therefore expand the coding potential of a genome. Cases are known of events that produce splicing variants (isoforms) showing novel and sometimes unexpected structural and functional properties [5,6]. On the other hand, evidence from analysis of sequences, structures and homology models suggest that many AS isoforms, even if detectable at the transcriptomic level, might not encode functional proteins because, for example, they lack important functional regions and/or seem to correspond to incomplete structures [7,8].

The overwhelming majority of AS evidence is based on transcriptomic data; therefore, a proof that the splicing product is eventually translated and can fold into a functional protein is generally missing. Nonetheless, it is evident that knowing whether or not an isoform observed at the transcriptional level does indeed correspond to a functional protein is relevant for both theoretical and practical reasons. Since it is practically impossible to identify negative cases - examples where one isoform certainly does not correspond to a functional protein - this is a scenario where we can only resort to computational methods for obtaining a probabilistic estimate of the likelihood that a protein is functional.

Computational method inferences are difficult to validate in the absence of a clearly defined negative set, but one can still assess their sensitivity in identifying

* Correspondence: anna.tramontano@uniroma1.it
[1]Dipartimento di Scienze Biochimiche, Sapienza Università di Roma, P.le A. Moro, 5 - 00185 Rome, Italy
Full list of author information is available at the end of the article

isoforms that are known to be translated because, for example, they have been unambiguously identified in proteomic experiments. Although the detection of a peptide identifying an isoform is not conclusive for its functional characterization, it does imply that the corresponding transcript is translated into a protein likely to fold and be produced at sufficient levels to be detected, and therefore strongly suggests that it is unlikely to be non-protein coding. This concept has been applied in the past, in small scale, to data by Tanner and coworkers [9], who found 16 human genes for which two different isoforms could be unambiguously identified by mass spectrometry (MS). A larger scale systematic analysis of isoform proteomic identification based on MS data performed for the fruit fly [10] led to the identification of AS events that could be confirmed at the protein level for 130 genes. The limited coverage of proteomics data, still far from the level of completeness provided by transcript expression analysis platforms [10], is the main reason behind the relatively low number of genes identified in both the aforementioned studies.

In this work, we take advantage of MS data for constructing a dataset composed of human isoforms unambiguously identified by MS (AS positive (ASPos) dataset) and use several computational methods to compare their properties with those of isoforms for which no matching peptide can be found in MS public database (unknown dataset). In particular, we study: their structural plausibility, based on structural models by homology; the presence of complete domains, based on Pfam domain definitions [11]; and the presence of functional sites, such as catalytic sites, based on SwissProt annotated features [12].

The results obtained with this positive dataset, which we used as a benchmark, allowed us to estimate how much each of the methodologies listed above can help in identifying translated isoforms. There is clearly a trade-off between the coverage achieved by each method (for example, the presence of a functional domain is more frequent than the presence of annotated functional sites) and their reliability in predicting the likelihood that the isoform is translated into a product. We used our positive set to estimate the fraction of false negatives detected by each method separately and by their combinations. In order to validate our conclusions, we also built two additional datasets, one containing ORFs obtained from the translation of non-coding transcripts (negative dataset) and one including all products of genes not undergoing AS and for which experimental evidence is available by MS (the noASPos dataset). The first dataset is somewhat artificial since its elements are only selected on the basis of the absence of termination codons in a sufficiently long ORF (at least 100 amino acids long) and is not really representative of realistic cases. On the other hand, the unknown dataset might contain isoforms that are not observed because they are only present at specific times or in specific cell types and isoforms that are not detected for technical reason by MS. This notwithstanding, the analysis of both datasets can be used to obtain an estimate of the false positive rate of the computational techniques.

Results obtained by considering the ASPos and unknown datasets show that, as expected, the single method with highest sensitivity - that is, the ability to correctly identify translated products - can be achieved by relying on the conservation of features annotated in SwissProt in the isoform, but this is not very frequent (coverage of about 14%), while a reasonably high coverage (81%), a good sensitivity (95%) and a specificity above 40% can be achieved by analyzing the presence of non-truncated Pfam domains. Combining the latter with an assessment of the plausibility of the modeled structure of the isoform increases the coverage by another 8%, with a decrease in sensitivity, but in this case the lower estimate for the specificity increases by at least 2%.

When the artificial non-coding dataset is used as a negative set, the results do not change substantially. Clearly, no SwissProt annotations exist for these transcripts, the presence of non-truncated Pfam domains still has the highest coverage (81%) and sensitivity (95%), with a specificity of around 30%. Also, the combination of structural plausibility and the presence of non-truncated Pfam domains produces a similar picture with coverage, sensitivity and specificity values of 87%, 93% and 33%, respectively.

A different balance between specificity and sensitivity can be required in different cases; therefore, we think that the results reported here can provide a useful guide to prioritizing experiments for different purposes.

Results and discussion

Proteomics technology can provide experimental evidence that a specific isoform is expressed, translated and sufficiently stable to be detected *in vivo*, although, unfortunately, it cannot be used to exclude the presence of a protein, nor can any other experimental technique provide such information. Nevertheless, the analysis of proteomic datasets can offer a repertoire of isoforms whose products are certainly present in the cell. Proteomics experiments provide a large amount of data, which are available in specialized databases such as PeptideAtlas [13] in the form of peptides with unambiguous mapping to protein sequences. In order to be able to detect the presence of a specific isoform in the midst of the whole spectrum of possible products of a gene, we first need to identify those isoform regions that are specific for one isoform, that is, that do not map to any other isoform of the same gene [14].

Of the 22,320 Ensembl57 [15] protein coding genes, 15,914 produce more than one isoform, and are therefore subject to AS. We did not include in this dataset those isoforms annotated as non-protein coding by Ensembl, and those differing only in their UTRs at the 5' or 3' end (therefore having identical coding regions), and ended up with 60,568 isoforms. In this group of alternative transcripts, we identified all regions (whole exons or exon portions) of each gene that are included in only one isoform (Figure 1). The detection of peptides mapping to such specific regions in MS experiments allows the unambiguous identification of the translation of the corresponding transcripts. PeptideAtlas human build peptides (May 2010) were mapped to the exons of these isoforms and classified as specific or unspecific accordingly. A total of 1,124 isoforms (from 1,025 genes) are identified by at least one specific peptide, and represent the set of isoforms whose existence is confirmed at the protein level. This figure is somewhat different from that reported in [14], where specific transcripts for 3,059 human alternatively spliced genes were identified using PeptideAtlas peptides, but this was expected since we used a more up-to-date release of PeptideAtlas in which the peptide mapping criteria were more stringent.

We focused our analysis on those genes having at least one isoform unequivocally identified by PeptideAtlas peptides and at least one other isoform for which no peptide mapping to its specific regions was found and that were predicted by PeptideSieve [16] to be detectable by the most popular current MS technologies (this being due to their charge, hydrophobicity, mass, secondary structure, and so on). When more than one ASPos or unknown isoform were present for a gene, we selected only the shortest and longest ones, respectively. We also verified that the results would not be affected if we were to use isoforms identified by at least two or more peptides (data not shown).

We also built a dataset containing the products of all genes that do not undergo AS and that are identified by at least one peptide present in PeptideAtlas and a dataset built by translating ORFs present in processed transcripts annotated as non-coding in Ensembl (see Materials and methods).

In conclusion, our datasets include 555 isoforms identified by MS (ASPos dataset), 555 isoforms corresponding to the same genes but for which no specific peptide is present in PeptideAtlas (unknown dataset), 865 products of genes that do not undergo AS (noASPos dataset) and 555 translated sequences from non-coding transcripts (negative dataset).

Our unknown dataset doubtlessly includes isoforms whose product is not detected since it is present only in specific tissues, cell cycle phases, developmental stages, or in the presence of specific stimuli, but a certain fraction can produce non-protein coding transcripts.

Our aim is to determine how many of the isoforms in the ASPos dataset that are identified as true products of a regulated AS event can be detected by different computational methods in order to evaluate their sensitivity. For the reason described above, the unknown set can only be used to estimate the lower limit of the specificity of the methods, while the negative dataset does not suffer this problem but is less representative of a real situation, even though we obtained the sequence by translating processed transcripts rather than random genomic sequences.

Positive isoforms are predicted to be structurally more plausible than unknown isoforms

Arguably, a considerable amount of non-functional AS will lead to polypeptide sequences that can not fold in a

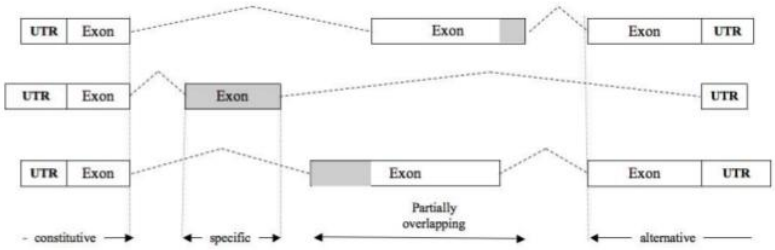


Figure 1 Scheme of possible scenarios for comparing different isoforms. Only peptides mapping in the products of shaded regions are considered specific.

stable conformation and therefore are quickly degraded. We cannot exclude that a stable conformation can be the result of profound structural rearrangements or of the establishment of stabilizing interactions with a binding partner. Such cases are very hard to identify, but apparently not very frequent [5,6].

Here we predicted the structure of all the isoforms of the ASPos and noASPos datasets for which a suitable structural template could be found and carefully analyzed the resulting models according to several criteria. We estimated how well packed the protein model is and the extent to which hydrophobic surface is exposed to solvent with respect to the average single domain proteins in the database of solved protein structures and to the template used to build the model (see Materials and methods). We also assessed whether insertions and deletions with respect to the structural template, when present, could be accommodated within the modeled structure. In particular, we flagged as 'unlikely' cases where a deletion would imply connectivity between two residues that are too far away in space and where insertions would occur in the well packed core of the protein.

The detailed pipeline for model building is described in the Materials and methods section. We were able to model 230 isoforms from the noASPos dataset, 147 from the ASPos dataset, 145 from the unknown dataset and 84 from the negative dataset, with coverage (that is, the fraction of protein sequence that can be modeled) of at least 90%.

The majority (134; 91%) of modeled isoforms from the ASPos dataset are structurally consistent. Difficult to accommodate deletions and/or insertions with respect to the template are present for nine isoforms from the positive dataset (6%), while five show a non-optimal packing of their interior (the two cases can obviously occur in the same isoform). The corresponding numbers for the noASPos dataset are similar (88% with a plausible structure, 9 isoforms with difficult to accommodate deletions/insertions corresponding to about 4% of the total and 18 with non-optimal packing corresponding to 8% of the total). On the other hand, the fraction of viable models is remarkably smaller for the negative (40; 48%) and unknown dataset isoforms (69; 48%). The negative dataset includes 13 models with difficult to accommodate deletions/insertions and 32 models with non-optimal packing. In 64 cases the models of the unknown dataset show non-optimal packing and in 10 they also have difficult to accommodate deletions/insertions.

Functional domains are more often truncated in unknown isoforms than in positive ones

AS can remove whole protein domains, but tend not to occur within domains [17]. While this is not an absolute rule, it is reasonable to assume that a substantial amount of isoforms where at least one domain is truncated by a splicing event correspond to non-protein coding transcripts.

To verify how well this criterion performs in real cases, we used the definition of domains in the Pfam database [11]. Each Pfam domain is described by a hidden Markov model (HMM) built on the seed example sequences for that domain. Different isoforms of the same gene can carry different sets of domains [18,19]. On the other hand, a domain that is truncated in an isoform is more likely to be the result of incorrect splicing than of a regulated event. As described in Materials and methods, a domain is considered truncated if the isoform sequence matches less than 70% of its length.

Most isoforms of the noASPos and ASPos datasets (83% and 86%, respectively) only include complete Pfam domains, and only 5% contain truncated Pfam domains. The situation is drastically different for the unknown and negative datasets, where 41% and 50% of the isoforms only contain complete Pfam domains, respectively, and 42% and 36% include at least one truncated Pfam domain.

It should be mentioned that more Pfam domains are found in isoforms of the ASPos dataset than the unknown one (2.43 domains on average versus 1.63). This could be attributed to the fact that these isoforms tend to be longer than the unknown ones (average length 694 amino acids versus 345). On the other hand, our data indicate that the length has little or no impact on the number of truncated domains. The average length of proteins included in the PDB database [20] is 606 amino acids and the percentage of proteins with truncated Pfam domains is only 14%, much lower than what we observe in our unknown dataset, and the percentage of truncated Pfam domains in these proteins is independent of the length (data not shown). Similarly, sequences in our noASPos dataset, whose average length (436) is comparable with that of members of the unknown (345) and negative datasets (485), include a truncated domain in only 5% of cases (compared with 42% and 36% for the unknown and negative datasets). Although we cannot exclude that the length of the transcripts might affect the results to a minor extent, we believe that in any case, the presence of a truncated domain, whatever the reason, is an indication of a lack or impairment of the associated function.

Functional features are rarely disrupted in positive isoforms

We verified whether AS would remove existing annotated active sites present in other isoforms of the same gene in the ASPos and unknown datasets. A similar procedure cannot be applied to the negative dataset,

since these translated sequences are not annotated, nor, obviously, to the noASPos dataset, where no AS occurs.

In several entries of our ASPos dataset, the annotation for an active site (80 genes) is present in the Swiss-Prot database [12]. Active sites are present in both isoforms of the positive and unknown dataset in only 60% of cases. In all other cases, the isoform of the unknown dataset (which is not identified by a specific peptide in PeptideAtlas) does not retain the active sites. In a single case a functional site is found in the unknown dataset isoform (Ensembl ID: ENSP00000359932) but not in the associated positive isoform (ENSP00000359935). In this case, positive and unknown isoforms have a radically different amino acid sequence after residue 114, due to the usage of different exons after the initial shared portion (for this reason these two isoforms are associated with different SwissProt IDs, FPGT_HUMAN and TNI3K_HUMAN, respectively) and have a different biological function (the former is a fucose-1-phosphate guanylyltransferase, the latter a serine/threonine-protein kinase). Therefore, in this particular case, the loss of functional sites is due to a radical functional change.

Transcription levels of the isoforms in different datasets

Isoform-specific expression can be estimated by means of recently developed microarray platforms that target all exons of a gene or (a subset of) exon-exon junctions. The Affymetrix Exon arrays [21] are high-density chips in which probe sets (composed by at least four probes) were designed for all exons in Ensembl. This platform proved to be very accurate and sensitive in the detection of AS events and is routinely used for the study of cellular processes in healthy or disease conditions. Expression data from 11 adult human tissues are publicly available from the Affymetrix website, and offer a valuable resource for the study of exon-level expression of human isoforms. The distribution of the expression level of the transcripts present in the ASPos, noASPos and unknown datasets are shown in Figure 2.

As can be appreciated from Figure 2, the isoforms in the unknown dataset have a lower level of normalized expression of specific exons than those in the ASPos and noASPos datasets. This notwithstanding, the overlap between the distributions is rather high (around 70%). Some of the unknown isoforms whose transcripts are expressed at lower levels might correspond to products present in limited amounts and less likely to be detected by MS, or they might be due to splicing errors, which are expected to happen at low frequency. On the other hand, a high percentage of non-detected isoforms are expressed at levels similar to those of the positive datasets and the lack of their detection points to the possibility that they do not produce functional proteins.

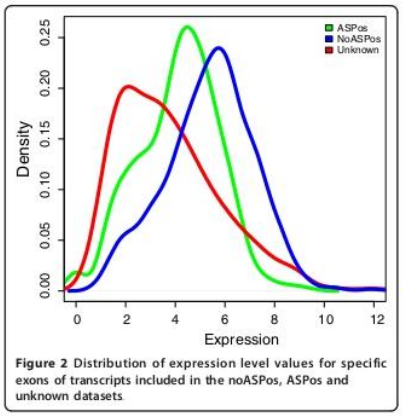


Figure 2 Distribution of expression level values for specific exons of transcripts included in the noASPos, ASPos and unknown datasets.

Statistical significance of different criteria

Figure 3 and Table 1 summarize results on coverage, accuracy, sensitivity and the lower estimate of the specificity of the criteria described above as well as of their union and intersection.

There is an obvious trade-off between coverage and accuracy. Isoforms that preserve the active sites, that contain only non-truncated Pfam domains and that are structurally plausible are very likely to be translated in functional products (80% accuracy), although this combination of features is only observed in a small fraction of the cases. On the other hand, a very good compromise for predicting the functionality of an isoform is to verify that it does not contain interrupted Pfam domains or has a plausible modeled structure. This would be appropriate for most practical purposes and applicable to almost 90% of the cases, providing an accuracy above 70% with a sensitivity and specificity of 93% and of at least 44%, respectively. The detection of complete Pfam domains, certainly easier to obtain in large scale analyses, has a high coverage and good sensitivity, although its specificity is not very high. In practice, when an isoform contains an interrupted Pfam domain, it is very likely not to be functional, while the detection of only complete Pfam domains in an isoform, especially if produced by a transcript overlapping with a coding one, as is the case for our non-coding dataset, is not very informative. The overall picture does not change when the non-coding dataset is considered instead of the unknown one (in which case, however, the conservation of the active sites cannot be taken into account), as shown in Table 2.

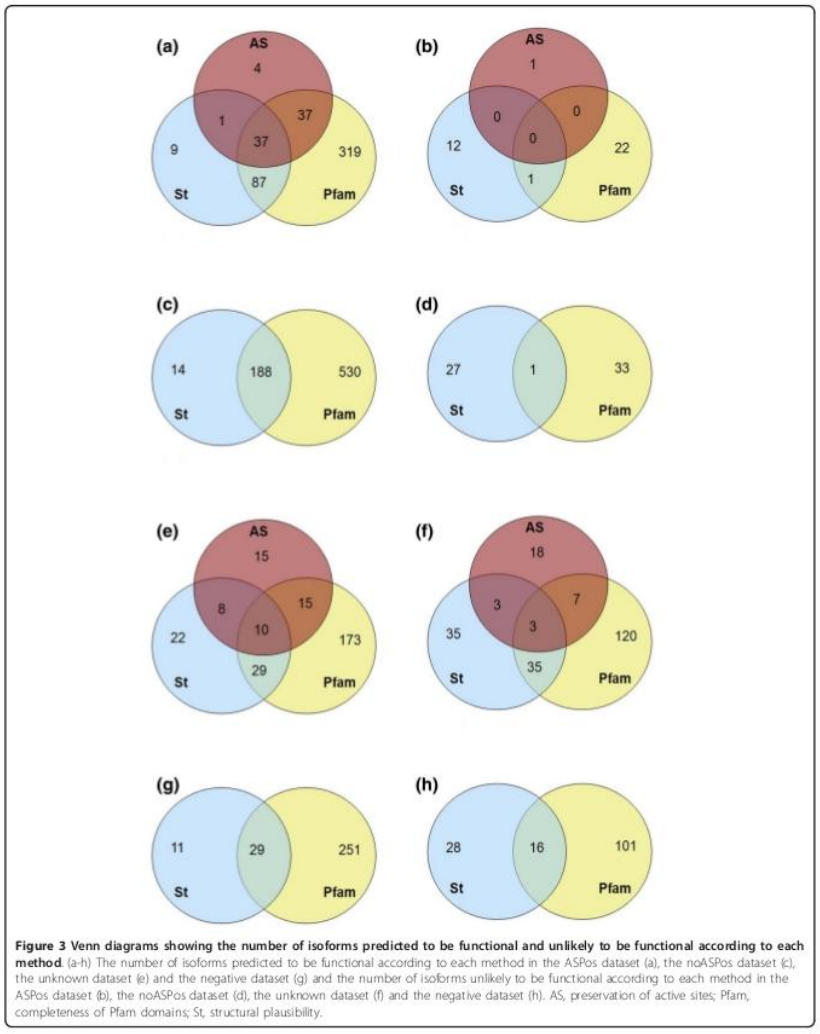


Figure 3 Venn diagrams showing the number of isoforms predicted to be functional and unlikely to be functional according to each method. (a-h) The number of isoforms predicted to be functional according to each method in the ASPos dataset (a), the noASPos dataset (c), the unknown dataset (e) and the negative dataset (g) and the number of isoforms unlikely to be functional according to each method in the ASPos dataset (b), the noASPos dataset (d), the unknown dataset (f) and the negative dataset (h). AS, preservation of active sites; Pfam, completeness of Pfam domains; St, structural plausibility.

Table 1 Results of the statistical analysis with respect to the unknown dataset

	Coverage	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity
AS	0.14	79	31	48	1	0.69	0.99	0.39
St	0.26	134	76	69	13	0.72	0.91	0.52
Pfam	0.81	480	165	227	23	0.72	0.95	0.41
AS ∪ St	0.35	175	101	99	14	0.71	0.93	0.50
St ∪ Pfam	0.89	490	203	257	35	0.71	0.93	0.44
AS ∪ Pfam	0.85	485	186	250	24	0.71	0.95	0.43
AS ∪ St ∪ Pfam	0.92	494	221	272	36	0.70	0.93	0.45
AS ∩ St	0.06	38	6	18	0	0.71	1.00	0.25
St ∩ Pfam	0.18	124	38	39	1	0.80	0.99	0.49
AS ∩ Pfam	0.10	74	10	25	0	0.77	1.00	0.28
AS ∩ St ∩ Pfam	0.05	37	3	10	0	0.80	1.00	0.23

Coverage, accuracy, sensitivity and specificity of the different strategies and their combinations (∪ = union and ∩ = intersection) with respect to the unknown dataset. AS, preservation of active sites; Pfam, completeness of Pfam domains; St, structural plausibility. The definition of the other parameters is reported in Materials and methods. FN, false negative; FP, false positive; TN, true negative; TP, true positive.

Conclusions

The wealth of high throughput data that are continuously being produced opens the way to the investigation of relevant properties of living organisms and can be effectively exploited in many instances. Cataloguing all putative isoforms of genes is one such example, although care should be taken since there is evidence that not all isoforms identified at the transcriptional level correspond to functional proteins [8].

The question that we address here - whether or not an isoform is likely to be functional - is relevant but unfortunately cannot be answered in a definite way by experimental approaches. While the presence of a functional protein in the cell can be demonstrated, it is impossible to assess that a given peptide sequence is not present or functional at any given time or in any compartment of a cell or an organism. Computational methods, provided they are properly assessed and evaluated, are therefore essential. We show here that different computational strategies and their combination can be effectively used as proxies for assessing the likelihood that an isoform observed at the transcriptional level does correspond to a functional protein product.

We believe that the estimate of the accuracy of different computational strategies and of their different combination provided here can be used for selecting different strategies for different occasions. In some cases, a higher rate of false positives might be preferable to a higher number of false negatives - for example, when a specific gene of interest is being investigated thoroughly - although even in this case prioritizing the experiment taking advantage of computational estimates can save time and resources.

Obviously, the impossibility of obtaining a true negative set implies that, while one can assess the ability of the methods to detect translated isoforms - that is, the percent of true positives and false negatives that they predict - it is impossible at present to give a precise estimate of how many false positives would result from any computational analysis. This is a very difficult, or perhaps impossible, problem to solve, but learning about the ability of the analyzed strategies to detect most of the truly translated isoforms and the lower estimate of their specificity that we have provided here can be of great help in understanding the functional repertoire of higher eukaryote genomes. Clearly, the accumulation of more and more proteomic data will allow even more effective strategies to be devised.

Materials and methods

The datasets used in this analysis were all constructed starting from the coding portion of the human genome in Ensembl57 [15]. Out of the total number of Ensembl protein coding genes (22,320), 6,406 genes are not

Table 2 Results of the statistical analysis with respect to the negative dataset

	Coverage	TP	TN	FP	FN	Accuracy	Sensitivity	Specificity
St	0.21	134	44	40	13	0.77	0.91	0.52
Pfam	0.81	480	117	280	23	0.67	0.95	0.29
St ∪ Pfam	0.87	490	145	291	35	0.66	0.93	0.33
St ∩ Pfam	0.15	124	16	29	1	0.82	0.99	0.36

Coverage, accuracy, sensitivity and specificity of the different strategies and their combinations (∪ = union and ∩ = intersection) with respect to the negative dataset. Pfam, completeness of Pfam domains; St, structural plausibility. The definition of the other parameters is reported in Materials and methods. FN, false negative; FP, false positive; TN, true negative; TP, true positive.

subjected to AS. Of all the isoforms encoded by the remaining genes, 31,618 are classified as non-protein coding according to the Ensembl annotations. In the remaining genes, we found 7,467 isoforms differing only for their untranslated regulatory regions from other isoforms in the same gene, and these were removed. We also discarded an additional 1,844 genes that were left with only one isoform. At this stage the dataset of alternative spliced isoforms contains 60,568 isoforms encoded by 13,980 genes.

Taken together, these latter isoforms contain 278,155 exons in their coding sequences identified by a unique Ensembl exon ID. These exons were classified as present in all transcripts of a gene (constitutive, 20% of the total), in a subset of the gene transcripts (semi-constitutive, 49%), or in a single transcript only (specific, 12%). Cases of semi-constitutive exons with parts of their sequence partially overlapping with another exon were classified in a separated category (partially overlapping, 19%).

Mapping of proteomic peptides on the human AS isoforms

Proteomics data were retrieved from the PeptideAtlas database [13]. PeptideAtlas organizes its data into builds centered on a particular species or tissue. We used the May 2010 human build [22], which contains 71,303 different peptides ranging in size from 7 to 66 (mean 17); these were unambiguously mapped to human Ensembl57 proteins. We selected only those peptides classified by PeptideAtlas as non-exon spanning. Of these, 39,956 match isoforms included in our dataset. We classified 11,005 peptides that unambiguously identify one protein isoform by mapping to a specific exon or to a specific part of partially overlapping exons as 'specific' peptides. Peptides that map into semi-constitutive exons, constitutive exons or non-specific parts of partially overlapping exons were classified as 'unspecific' peptides.

Building of the positive, negative and unknown datasets

The noAS positive dataset was built by selecting the products of all non-alternatively spliced genes that are unambiguously identified by PeptideAtlas peptides and contains 865 gene products identified by 4,589 peptides. All the isoforms produced by AS that are unambiguously identified by specific PeptideAtlas peptides (576 isoforms identified by 2,546 peptides) were considered for inclusion in the ASPos dataset. Out of all remaining isoforms, those having specific exons (or specific exon regions in partially overlapping exons) but that were not identified by any PeptideAtlas peptide, although they could in principle be detected according to the PeptideSieve algorithm [16], were considered for inclusion in the unknown dataset (782 isoforms). In detail, the sequences of the isoforms in the unknown dataset were submitted to the PeptideSieve algorithm, which predicts the likelihood of the peptide being observed in a proteomics experiment, taking into account ionization and missed cleavage propensity. The program first performs an *in silico* digestion of the protein and then computes for each peptide a list of physical and chemical descriptors. Next, it scores the likelihood that each peptide is observed in one of the four proteomics platforms (PAGE MALDI, PAGE ESI, ICAT ESI, MUDPIT ESI). An unknown isoform is considered detectable in a proteomics experiment if at least one of its peptides, originating from its specific regions, has a score of at least 0.5 (the default lower limit score in PeptideSieve). When used as described above, PeptideSieve has an expected accuracy above 85%. When more than one identified or not identified isoform was present in the same gene, we included only the shortest one of the ASPos dataset and the longest one of the unknown dataset. At the end of the procedure the ASPos and unknown datasets included 555 isoforms each.

To obtain the negative dataset, we considered transcripts annotated as non-coding present in regions of the genome containing at least one coding transcript (we did not include transcripts undergoing nonsense mediated decay), translated their sequence starting from the first AUG and continuing until a stop codon was encountered, and selected the longest 555 translated sequences. The average lengths of members of the datasets are 436 (noASPos), 694 (ASPos), 345 (unknown) and 485 (negative).

Structural characterization of the isoform datasets

We built structural models by homology for each isoform in our datasets for which the native structure is unknown, and for which a suitable template covering more than 90% of the sequence could be found. HHsearch 1.1.5 [23] was used to search for possible structural templates (default parameters) and for obtaining the sequence alignment between the target and its putative templates. The resource builds a HMM of the target protein family and compares it to the HMMs representing a set of non-redundant families of proteins of known structure (sequence identity between any pair below 70%). Model building was performed using a local version of Modeller9v8 [24] (default parameters). Models were considered structurally plausible if there is a deletion with respect to the template and the distance between the two residues on either side is larger than 15Å; if there is an insertion of more than three residues in the core of the protein, that is, between two residues whose solvent accessibility calculated with POPS [25] is lower than $5Å^2$; if the packing efficiency of the resulting model computed using the OS software [26] is below

0.54 while that of the template used for modeling is not; and if the 'packing-eff' computed using the NUCPROT package [27] is below 25.9, while that of the template used for modeling is not. The thresholds for POPS, Packing-eff and OS tools were derived by running the programs on 4,122 monomeric proteins solved by X-ray crystallography at a resolution better than 2Å. The chosen thresholds, 25.5 for POPS values, 25.9 for Packing-eff values and 0.54 for OS values, correspond to two standard deviations from the average (data not shown).

Functional domain characterization of the isoform datasets

We mapped Pfam domains [11] on the protein sequences of the isoforms in the datasets, using the batch search utility available through the Pfam web interface, using Pfam-A families and an E-value below 10E-5. For each domain, we computed the coverage of the HMM representing the domain: all domains assigned to an isoform whose length covers less than 70% of the corresponding HMM were considered truncated. This threshold was chosen by evaluating the HMM coverage in a set of protein sequences for which the structure is known. We extracted 3,859 monomeric structures with less than 30% sequence identity from PDB [20]. For every protein, the corresponding Uniprot sequence was retrieved and Pfam domains were assigned according to the criteria described above; 93% of Pfam domains have a coverage between 0.70 and 1.0 in these sequences (data not shown).

Mapping of Swiss-Prot features on the isoform datasets

Active site residues as annotated in Swiss-Prot (release 57, March 2009) were mapped on all isoforms encoded by a gene using the Ensembl *perl* APIs and using in-house developed tools.

Evaluation of transcriptomic expression

Exon-level isoform expression was extracted from Affymetrix Exon 1.0 ST Array public datasets [28]. The profiled human tissues include breast, cerebellum, heart, kidney, liver, muscle, pancreas, prostate, spleen, testis, and thyroid. RNA-normalized probe-set expression levels, computed using the Affymetrix Power Tools (APT), are available from the Affymetrix web site. We were able to retrieve isoform-specific expression levels for 728 and 532 isoforms of the noASPos and ASPos datasets, respectively, and 264 isoforms of the unknown dataset. Probe set expression levels were computed as the median of normalized expression values in the 11 tissues in the panel. Isoform expression is estimated as the median expression of all probe sets falling in the isoform-specific regions.

Data analysis

We used the following definitions: true positive (TP) - an isoform in the positive dataset for which the considered descriptor is consistent with the hypothesis of the isoform being functional (that is, structurally plausible, or not containing truncated domains, or containing an active or binding site); false negative (FN) - a positive set isoform for which a descriptor suggests loss of functionality (that is, structurally not plausible, or containing a truncated domain, or missing active sites present in some other isoform of the same gene); false positive (FP) - an unknown or negative set isoform for which the considered descriptor is consistent with the hypothesis of the isoform being functional; true negative (TN) - an unknown or negative set isoform for which a descriptor suggests loss of functionality. We can use the confusion matrix originated by the values of TP, FN, FP, TN to evaluate how well each descriptor (and all their intersections and unions) is able to discriminate between isoforms in the datasets, using the commonly used measures accuracy (ratio between correct predictions TP + TN and total predictions TP + FN + FP + TN), sensitivity (ratio between correct positive predictions TP and total predictions in the positive dataset TP + FN), and specificity (ratio between correct negative predictions TN and total predictions in the negative or unknown dataset TN + FP). Since each descriptor can be applied only to a subset of the total isoforms (for example, not all isoforms can be modeled, not all isoforms have Pfam domains, and not all isoforms have an annotated active site), the coverage of each descriptor (defined as the number of isoforms to which the descriptor can be applied over the total number of isoforms under examination) is highly variable. The union of two descriptors, or of all three of them, obviously increases the coverage at the cost, in some cases, of accuracy. When considering the union of two or three descriptors, one must take into account that a number of isoforms can have discordant descriptors. For example, a given isoform of the positive dataset can be structurally plausible, thus having one truncated domain, and therefore it is a TP from the point of view of the structural descriptor ad a FN for the domain integrity descriptor. In all these cases we counted these isoforms as FN (or as TN for isoforms belonging to the negative or unknown dataset).

Abbreviations

AS: alternative splicing; ASPos: alternatively spliced positives; FN: false negative; FP: false positive; HMM: hidden Markov model; MS: mass spectrometry; noASPos: non alternatively spliced positives; ORF: open reading frame; TN: true negative; TP: true positive; UTR: untranslated region.

Acknowledgements

This work was supported by award number KUK-I1-012-43 made by King Abdullah University of Science and Technology (KAUST), by FIRB Italbionet

and Proteomica and by Ministero della Salute Progetto RF-ID-2006-354931. The authors are grateful to Matteo Floris for useful discussions.

Author details
[1]Dipartimento di Scienze Biochimiche, Sapienza Università di Roma, P.le A. Moro, 5 - 00185 Rome, Italy. [2]INRAN, Via Ardeatina, 546 - 00178 Roma, Italy. [3]Istituto Pasteur Fondazione Cenci Bolognetti, Sapienza Università di Roma, P.le A. Moro, 5 - 00185 Rome, Italy.

Authors' contributions
GL, LL, FF and RD carried out the analysis and helped to draft the manuscript. AT conceived of the study, participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 31 July 2010 **Revised:** 17 December 2010
Accepted: 20 January 2011 **Published:** 20 January 2011

References

1. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**:621-628.
2. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML: **A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.** *Science* 2008, **321**:956-960.
3. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes.** *Nature* 2008, **456**:470-476.
4. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.** *Nat Genet* 2008, **40**:1413-1415.
5. Birzele F, Csaba G, Zimmer R: **Alternative splicing and protein structure evolution.** *Nucleic Acids Res* 2008, **36**:550-558.
6. Stetefeld J, Ruegg MA: **Structural and functional diversity generated by alternative mRNA splicing.** *Trends Biochem Sci* 2005, **30**:515-521.
7. Melamud E, Moult J: **Structural implication of splicing stochastics.** *Nucleic Acids Res* 2009, **37**:4862-4872.
8. Tress ML, Martelli PL, Frankish A, Reeves GA, Wesselink JJ, Yeats C, Olason PI, Albrecht M, Hegyi H, Giorgetti A, Raimondo D, Lagarde J, Laskowski RA, López G, Sadowski MI, Watson JD, Fariselli P, Rossi I, Nagy A, Kai W, Størling Z, Orsini M, Assenov Y, Blankenburg H, Huthmacher C, Ramirez F, Schlicker A, Denoeud F, Jones P, Kerrien S, *et al*: **The implications of alternative splicing in the ENCODE protein complement.** *Proc Natl Acad Sci USA* 2007, **104**:5495-5500.
9. Tanner S, Shen Z, Ng J, Florea L, Guigo R, Briggs SP, Bafna V: **Improving gene annotation using peptide mass spectrometry.** *Genome Res* 2007, **17**:231-239.
10. Tress ML, Bodenmiller B, Aebersold R, Valencia A: **Proteomics studies confirm the presence of alternative protein isoforms on a large scale.** *Genome Biol* 2008, **9**:R162.
11. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2010, , **38 Database:** D211-222.
12. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31**:365-370.
13. Deutsch EW, Lam H, Aebersold R: **PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows.** *EMBO Rep* 2008, **9**:429-434.
14. Blakeley P, Siepen JA, Lawless C, Hubbard SJ: **Investigating protein isoforms via proteomics: a feasibility study.** *Proteomics* 2009, **10**:1127-1140.
15. Flicek P, Aken BL, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Gräf S, Haider S, Hammond M, Howe K, Jenkinson A, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Koscielny G, Kulesha E, Lawson D, Longden I, Massingham T, McLaren W, *et al*: **Ensembl's 10th year.** *Nucleic Acids Res* , **38 Database:** D557-562.
16. Mallick P, Schirle M, Chen SS, Flory MR, Lee H, Martin D, Ranish J, Raught B, Schmitt R, Werner T, Kuster B, Aebersold R: **Computational prediction of proteotypic peptides for quantitative proteomics.** *Nat Biotechnol* 2007, **25**:125-131.
17. Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, Gelfand MS, Sunyaev S: **Increase of functional diversity by alternative splicing.** *Trends Genet* 2003, **19**:124-128.
18. Floris M, Orsini M, Thanaraj TA: **Splice-mediated Variants of Proteins (SpliVaP) - data and characterization of changes in signatures among protein isoforms due to alternative splicing.** *BMC Genomics* 2008, **9**:453.
19. Beaussart F, Weiner J, Bornberg-Bauer E: **Automated Improvement of Domain Annotations using context analysis of domain arrangements (AIDAN).** *Bioinformatics* 2007, **23**:1834-1836.
20. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28**:235-242.
21. Gardina PJ, Clark TA, Shimada B, Staples MK, Yang Q, Veitch J, Schweitzer A, Awad T, Sugnet C, Dee S, Davies C, Williams A, Turpaz Y: **Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array.** *BMC Genomics* 2006, **7**:325.
22. **PeptideAtlas May 2010 Human Build.** [http://www.peptideatlas.org/builds/human/201005/APD_Hs_all.fasta].
23. Söding J: **Protein homology detection by HMM-HMM comparison.** *Bioinformatics* 2005, **21**:951-960.
24. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A: **Comparative protein structure modeling using Modeller.** *Curr Protoc Bioinformatics* 2006, **Chapter 5**:Unit 5.6.
25. Cavallo L, Kleinjung J, Fraternali F: **POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level.** *Nucleic Acids Res* 2003, **31**:3364-3366.
26. Pattabiraman N, Ward KB, Fleming PJ: **Occluded molecular surface: analysis of protein packing.** *J Mol Recognit* 1995, **8**:334-344.
27. Voss NR, Gerstein M: **Calculation of standard atomic volumes for RNA and comparison with proteins: RNA is packed more tightly.** *J Mol Biol* 2005, **346**:477-492.
28. **Affymetrix Gene 1.0 ST Array Data Set.** [http://www.affymetrix.com/support/technical/sample_data/gene_1_0_array_data.affx].

doi:10.1186/gb-2011-12-1-r9
Cite this article as: Leoni *et al.*: Coding potential of the products of alternative splicing in human. *Genome Biology* 2011 **12**:R9.

Paper II: MAISTAS a tool for automatic structural evaluation of alternative splicing products

BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 12 2011, pages 1625–1629
doi:10.1093/bioinformatics/btr198

Structural bioinformatics Advance Access publication April 15, 2011

MAISTAS: a tool for automatic structural evaluation of alternative splicing products

Matteo Floris[1,†], Domenico Raimondo[2,†], Guido Leoni[2], Massimiliano Orsini[1], Paolo Marcatili[2] and Anna Tramontano[3,4,*]

[1]CRS4-Bioinformatics Laboratory, c/o Sardegna Ricerche Scientific Park, Pula, 09010 Cagliari, [2]Department of Biochemical Sciences, [3]Department of Physics and [4]Istituto Pasteur Fondazione Cenci Bolognetti, Sapienza University of Rome, P le Aldo Moro 5, 00185 Rome, Italy

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Analysis of the human genome revealed that the amount of transcribed sequence is an order of magnitude greater than the number of predicted and well-characterized genes. A sizeable fraction of these transcripts is related to alternatively spliced forms of known protein coding genes. Inspection of the alternatively spliced transcripts identified in the pilot phase of the ENCODE project has clearly shown that often their structure might substantially differ from that of other isoforms of the same gene, and therefore that they might perform unrelated functions, or that they might even not correspond to a functional protein. Identifying these cases is obviously relevant for the functional assignment of gene products and for the interpretation of the effect of variations in the corresponding proteins.

Results: Here we describe a publicly available tool that, given a gene or a protein, retrieves and analyses all its annotated isoforms, provides users with three-dimensional models of the isoform(s) of his/her interest whenever possible and automatically assesses whether homology derived structural models correspond to plausible structures. This information is clearly relevant. When the homology model of some isoforms of a gene does not seem structurally plausible, the implications are that either they assume a structure unrelated to that of the other isoforms of the same gene with presumably significant functional differences, or do not correspond to functional products. We provide indications that the second hypothesis is likely to be true for a substantial fraction of the cases.

Availability: http://maistas.bioinformatica.crs4.it/.

Contact: anna.tramontano@uniroma1.it

Received on October 26, 2010; revised on March 17, 2011; accepted on March 22, 2011

1 INTRODUCTION

Determining the identity and function of all the sequence elements in human DNA is a daunting challenge. The large scale pilot phase of the ENCODE project (Birney *et al.*, 2007) provided an exhaustive identification and verification of functional sequence elements in a limited region of 1% of the human genome. The computational analysis of the data revealed several unexpected features of the genome (Tress *et al.*, 2007). Perhaps the most surprising one was that many transcribed elements could be neutral elements that serve as a reservoir for natural selection. Many of these transcripts derive from alternative splicing events. Their putative products were manually analysed by the BioSapiens European Consortium (Tress *et al.*, 2007). The analysis led to the striking conclusion that more than 50% of them might not give rise to proteins structurally and/or functionally related to the other isoforms of the same genes or be the result of aberrant splicing events giving rise to non-functional proteins (Tress *et al.*, 2007).

Indeed, comparison of the putative proteins encoded by the alternatively spliced transcripts with the main isoform showed that most of them lacked an active site, key trans-membrane segments, essential signalling regions and post-transcriptionally modified sites. Most importantly, models of their putative three-dimensional structures did not seem to correspond to plausible folds (Tress *et al.*, 2007).

This observation was confirmed by Moult and co-workers (Melamud and Moult, 2009a, b) who, using a completely different dataset of alternative splicing variants, found that the vast majority of them resulted in putatively unstable protein conformations.

Recently, some of us manually analysed the putative structures of isoforms of the human genome, the existence of which had been confirmed by mass-spectrometry and of isoforms of the same genes for which no evidence exists in proteomic databases reaching essentially the same conclusions (Leoni *et al.*, 2011).

Altogether these observations suggest that we might be observing the effects of noisy selection of splice sites by the splicing machinery and/or that alternatively spliced products of a gene might assume unrelated conformations.

These findings raise several interesting questions, but also a few practical issues. First of all, the careful manual analysis performed by the BioSapiens consortium on 1% of the genome needs to be scaled up to the whole genome and therefore automated. Secondly, analysis tools should be available to biologists performing experiments in a user-friendly manner.

At present, there are a few systems that partially satisfy this need. For example, the ProSas database (Birzele *et al.*, 2008) (http://www.bio.ifi.lmu.de//forschung/structural-bioinformatics/prosas) stores structures and models (provided the target proteins shares at least 40% sequence identity with a known template) for the alternative isoforms annotated in Ensembl (Hubbard *et al.*, 2002)

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

and Swiss-Prot (Bairoch *et al.*, 2004) and allows the visualization of the exon boundaries in the context of the three-dimensional structures, but there is no provision for automatic analysis of the plausibility or completeness of the resulting structures and models. The same is true for AS-ALPS (Shionyu *et al.*, 2009) (http://as-alps.nagahama-i-bio.ac.jp/), a server that provides information about the putative effect of alternative splicing on human and mouse proteins, provided that at least one of the isoforms has an experimentally solved structure.

Here, we describe a system named Modelling and Assessment of ISoforms Through Automated Server (MAISTAS) that, given the accession codes of one or more genes or proteins, collects all their putative spliced isoforms annotated in the Ensembl genome database (Hubbard *et al.*, 2002), builds, whenever possible, comparative models for their structures, analyses their features and provides an estimate of the likelihood that the isoforms correspond to potentially stable and structurally plausible proteins in the absence of major conformational rearrangements.

Alternative splicing isoforms can also be uploaded in the FASTA format in order to allow the user to analyse data from more comprehensive and specialized databases such as Aceview (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/) (Thierry-Mieg and Thierry-Mieg, 2006) or ASPicDB (http://t.caspur.it/ASPicDB/) (Martelli *et al.*, 2010).

Model assessment is performed by analysing the quality of the packing in the core of the structure and/or model, the extent of exposed hydrophobic surface and the putative effect of deletions and insertions. These properties are compared to those observed in known protein structures and in the closest homologs of the known structure. The system is freely available as a Web server.

2 METHODS

The input data can be a set of sequences in the FASTA format or one or more of the following codes: Ensembl Gene ID(s), Ensembl Transcript ID(s), Ensembl protein ID(s) (Flicek *et al.*), EMBL ID(s) (Leinonen *et al.*, 2011), EntrezGene ID(s) (Maglott *et al.*, 2011), GO ID(s) (Ashburner *et al.*, 2000), HGNC automatic gene name, HGNC curated gene name (Seal *et al.*, 2011), UniProt/TrEMBL Accession(s), UniProt/Swissprot ID(s), UniProt/Swissprot Accession(s) (The Uniprot Consortium, 2008), VEGA transcript ID(s), HAVANA transcript ID(s) (Wilming *et al.*, 2008).

The collection of all putative splicing isoforms corresponding to the input gene (or to the gene encoding for the protein when a protein accession code is used) is achieved by taking advantage of a locally stored version of the Ensembl database (release 58) (Flicek *et al.*, 2011). Users can select accession codes for more than 30 different organisms.

The HHsearch 1.1.5 (Söding, 2005) is used to search for possible structural templates (*E*-value lower than 10^{-5}, sequence coverage of at least 90%, global alignment mode, all other parameters set at their default values) and for obtaining the sequence alignment between the target and its templates. Model building is performed using a local version of Modeller9v8 (Sali and Blundell, 1993) (default parameters).

The selected parameters ensure that the quality of the produced models is sufficiently high to be able to reliably measure properties described below as demonstrated by the last CASP experiment (http://predictioncenter.org/CASP9).

POPS (Cavallo *et al.*, 2003) is used to calculate the accessibility to the solvent of each residue of the models. The OS software (Pattabiraman *et al.*, 1995; Fleming and Richards, 2000) is used for computing infrequent environment of residues. Finally, the 'packing-eff' method from the NUCPROT package (Voss and Gerstein, 2005) is used for estimating how well packed the protein is.

The thresholds for POPS, Packing-eff and OS tools were derived by running the programs on 7908 monomeric proteins solved by X-ray crystallography at a resolution better than 2.5 Å. The chosen thresholds, 20.1 for POPS values, 17.8% for Packing-eff values and 0.54 for OS values, correspond to two standard deviations from the average (data not shown).

Residues are considered exposed if their mean solvent accessibility—calculated considering three residues on each side of them—is larger than 5 $Å^2$.

The average response time for a typical request (three to four isoforms, a few hundreds amino acid long) is <1 h, the time limiting factor being the construction of the HMMs and of the corresponding models. The entire pipeline was built using python scripts and the interface is PHP based.

In order to verify that the system can be applied to a substantial fraction of cases and that is able to recognize translated proteins, we ran it on protein isoforms whose existence is unambiguously identified by mass spectrometry. We used the May 2010 human build (http://www.peptideatlas.org/builds/human/201005/APD_Hs_all.fasta) containing 72 396 different peptides ranging in size from 6 to 66 (mean 17) (Deutsch *et al.*, 2008). Of these, 19 513 could be unambiguously mapped to 2972 isoform products annotated in Ensembl (release 58). We also compared the results of MAISTAS with those obtained by a manual analysis of human transcript products described in Leoni *et al.* (2011).

3 RESULTS

The automatic analysis performed by MAISTAS requires that the user inputs one or more protein/gene accession codes from the common public databases (see Section 2) or a set of sequences in the FASTA format. In all but the last case, the sequence(s) corresponding to the user query is retrieved and mapped back to the appropriate genome database by using a local installation of the BioMart database (Durinck *et al.*, 2005). The peptide sequences of all isoforms of the target gene, as annotated in Ensembl, are retrieved.

If the input is a set of amino acid sequences in the FASTA format, they are assumed to be different isoforms of the same gene.

The user can supply an email address (optional) to which the results will be sent or bookmark the result page. The initial query page of MAISTAS provides a link to an example result page, which allows the user to inspect a typical output (Fig. 1).

In the first step, the tool evaluates whether a structure exists for any of the isoforms or, lacking this, whether a comparative model can be built. In the latter case, the template is identified using the HHsearch program, which builds a Hidden Markov Model (HMM) of the target protein family and compares it to the HMMs representing a set of non-redundant families of proteins of known structure (sequence identity between any pair below 70%). This strategy has been shown in blind tests to be one of the most sensitive for finding structural templates (Battey *et al.*, 2007).

The target sequence, the template(s) and the alignment obtained by the HHsearch are automatically analysed. Only models based on template structures solved by X-ray crystallography or an NMR are considered. They are inspected to detect any possible gaps in the coordinate set (for example, because of the absence of electron density in X-ray structures). If these regions are present at the N- or C-terminus of the protein they are trimmed, otherwise a warning is issued. A warning is also issued if the alignment includes insertions larger than 50 residues that might correspond to an inserted domain or deletions larger than 20 residues.

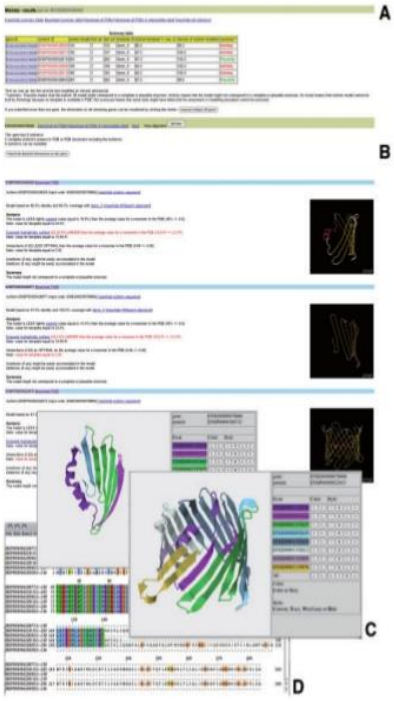


Fig. 1. Snapshots of the MAISTAS output page. (**A**) Summary table for the modelled isoforms. The following data are shown: gene ID (gene identification code), isoform ID (isoform identification code), isoform length (number of residues of each isoform), first aa, last aa (the first and last modelled or solved amino acid), template ID (the PDB code of the template protein used for modelling or the PDB code of the known isoform structure), isoform/template % seq. ID (sequence identity between the splicing isoform and the sequence of the selected template), fraction of isoform modelled (percentage of the splicing isoform sequence modelled), summary (assessment of the plausibility of the structure). (**B**) Snapshot of the isoform section showing results of the analysis for each isoform, its final assessment and the modelled structure in a small Jmol window. Different links in the section allow the user to download the coordinates of the model, view their 3D structure with regions corresponding to exons in different colours, view the amino acid sequence and the isoform/template alignment generated by the HHsearch. (**C**) Alternative spliced isoform three-dimensional structures are displayed in separate windows allowing their simultaneous analysis and comparison. On the right side of each Jmol window, the user can choose which exons should be displayed and select different representation modes. By default, all exons are mapped on the protein structure, each in a different colour. (**D**) Multiple sequence alignment of the isoforms displayed via the JALVIEW applet.

The alignment is used to build the model using a local installation of Modeller (Sali and Blundell, 1993). Once the model has been built, the system computes the model hydrophobic solvent accessible area and packing efficiency.

If the modelled isoform presents deletions with respect to the template, the Euclidean distance between the Cα residues before and after the deletion(s) is recorded. If insertions are present, the surface exposed to the solvent of the amino acids surrounding them and the number of inserted amino acids is computed.

The tool informs the user that the model might not correspond to a complete or plausible structure if the distance between the two residues on either side of a deletion is >15 Å and/or if there are more than three residues inserted in the core of the protein and/or if the hydrophobic solvent accessible area of the model is larger than a set threshold (see Section 2). In assessing the results, the system takes into account the corresponding values for the template used for modelling.

The output of MAISTAS is shown in Figure 1 and includes a summary table, where all the data regarding the modelled isoforms are reported. These can also be downloaded as a csv file. The user can download the coordinates of all the models and, if desired, all the intermediate data used in the procedure. The next section of the output page describes the detailed results for each modelled isoform and reports (see Section 2 for details):

- The sequence identity and coverage of the template and its PDB code.
- The packing efficiency of the model and of its template together with their comparison with the expected value.
- The extent of the exposed hydrophobic area of the model and of its template together with their comparison with the expected value.
- The packing environment of residues in the model and the template together with their comparison with the expected value.
- The assessment of whether insertions and deletions (if any) can be easily accommodated into the model.
- The modelled or experimental structure in a Jmol window.
- The option to inspect the multiple sequence alignment via a JALVIEW applet (Waterhouse *et al.*, 2009).
- The option to visualize and analyse the models via a Jmol applet (http://www.jmol.org/).
- A final remark about the plausibility/completeness of the predicted structure.

MAISTAS depends on the availability of structural templates to predict the three-dimensional structure of the isoforms by comparative modelling. If no structural templates are available, a 'No template satisfying all parameters' warning is issued. When MAISTAS is unable to provide a reasonable structural model (e.g. when very large insertions are present) the system will return the message 'Maistas is having trouble modelling or assessing this isoform'.

The online result pages are accessible via the URL sent either by e-mail or via the 'Retrieve results by job identifier or by email' window, using the provided job identification code or the e-mail address.

Produced models and the results of their analysis are stored in a local database unless the user requests them to be kept private. This implies that a user might be able to immediately retrieve the results on the gene(s) of interest if they were already been produced in a previous run of the system. The entries of the database are time stamped and presented to the user together with an option to repeat the analysis, which is advisable if major updates of the genome or structure database have taken place since the previous analysis was performed.

We ran the system on all human alternatively spliced isoform whose existence at the protein level could be unambiguously verified by mass spectrometry, i.e. of those protein isoforms for which a peptide that unambiguously identifies them has been detected with high reliability by mass spectrometry.

The server was able to produce and analyse models in 30% of the cases (890 out of 2972). In 2082 of them (70%), the model could not be built because there is no template satisfying all parameters. This had to be expected since we use rather stringent parameters to select the template (*E*-value better than 10^{-5}, template coverage >90%, X-ray resolution <2.5 Å or solved by the NMR). Out of the modelled isoforms, 712 (80%) were assessed as structurally plausible (see http://www.bioinformatica.crs4.org/maistas/pub/dataset.xls). In the majority of the remaining cases, (160 out of 178) the model showed a large hydrophobic surface exposed to the solvent. In these cases, the protein might indeed represent an incomplete and therefore not plausible structure, but also simply be a subunit of a larger complex.

We compared the results obtained by MAISTAS with those derived from a manual analysis of the isoforms of genes for which at least one isoform had been detected in mass-spectrometry experiments [and unambiguously identified by the presence of a peptide in the PeptideAtlas database (Deutsch *et al.*, 2008) and at least one had not (Leoni *et al.*, 2011)]. The results obtained automatically using MAISTAS are consistent with those reported in Leoni *et al.* (2011). In particular, MAISTAS was able to model 30% of the 555 proteins for which there is an evidence of translation (to be compared with the 26.4% obtained in the manual analysis), 85% of which were assessed as structurally plausible. The difference in coverage between the manual and automatic analyses is due to the increased size of the protein sequence and structure databases. Models were also produced for 181 out of 555 isoforms for which there is no evidence of translation in PeptideAtlas. Only 44% of these isoforms were reported as complete and plausible by the automatic pipeline. The corresponding numbers for manual analysis are 145 isoforms (26%) modelled and 48% classified as structurally consistent.

3.1 Application example

As an example of the use of MAISTAS, we describe the results obtained using the gene coding as input for the voltage-dependent anion channel 3 (VDAC3) (Ensembl gene identification code: ENSG00000078668), a protein that forms a channel through the mitochondrial outer membrane allowing diffusion of small hydrophilic molecules. Six splice variants are present in the Ensembl database for the gene encoding the protein, identified by the following Ensembl peptide codes: ENSP00000428845, ENSP00000022615, ENSP00000428519, ENSP00000428977, ENSP00000429006 and ENSP00000428029.

The UniProt database entry of VDAC3 (Q9Y277) describes only two of these isoforms (ENSP00000388732 and ENSP00000022615). Although four peptides mapping to the putative products are present in the PeptideAtlas database (PeptideAtlas IDs: PAp00006999; PAp00007806; PAp00077146; and PAp00423732), they cannot be used to unambiguous identify specific isoforms of the gene since they fall in the exons present in all of them.

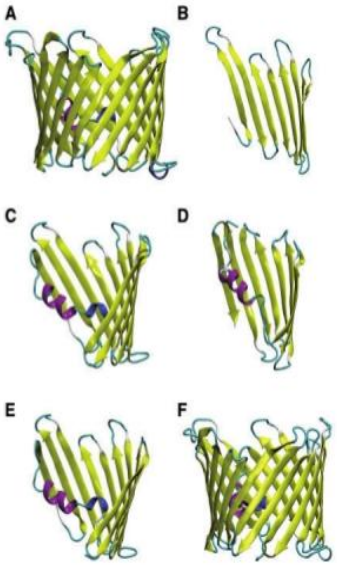


Fig. 2. Three-dimensional models of the VDAC3 protein isoforms. (**A**) ENSP00000428845. (**B**) ENSP00000428977. (**C**) ENSP00000428519. (**D**) ENSP00000428029. (**E**) ENSP00000429006. (**F**) ENSP00000422615.

Decker *et al.* (Decker and Craigen, 2000) used specific anti-VDAC3 antibody and demonstrated the existence of the ENSP00000428845 and ENSP00000022615 isoforms. The only difference between these two alternatively spliced isoforms is the insertion of a single methionine at position 39 of the ENSP00000428845 sequence.

ENSP00000022615 is also annotated in the CCDS database, a resource that centralizes the identification of well-supported, consistently annotated, protein-coding regions (Pruitt *et al.*, 2009). MAISTAS was able to provide a plausible structural model for isoforms ENSP00000428845 and ENSP00000022615 (Fig. 2A and F), while models of ENSP00000428519, ENSP00000428977, ENSP00000429006 and ENSP00000428029 were considered unlikely or incomplete (Fig. 2B–E). Inspection of the HHpred alignment used for building the ENSP00000428519, ENSP00000428977, ENSP00000429006 and ENSP00000428029 isoform models does not highlight any specific problem with the alignment (data not shown); however, the VDAC3 beta-barrel domain architecture is completely disrupted in the models of ENSP00000428519, ENSP00000428977, ENSP00000429006 and

ENSP00000428029 (Fig. 2B–E). All these isoforms show a large exposed hydrophobic surface, (around 22 Å^2, compared with the expected value of 15.6 Å^2 and with the value observed for the template of 15.9 Å^2). This dramatic architecture variation might imply that the isoforms are non-functional or that they perform a completely different function.

4 CONCLUSION

The more detailed is the analysis of the genomes of higher eukaryotes, the more complex they are revealed to be. For example, it is becoming clear that alternative splicing events do not simply result in a modulation of the function of the gene products, for example, by removing or adding structurally compact domains, or by modifying the sequence of specific regions of the encoded protein, but that they can either have a profound effect on the structure and function of the products of the same gene or give raise to non-functional products (Melamud and Moult, 2009a, b; Tress *et al.*, 2007).

The latter can nevertheless have a relevant biological function. For example, Poliseno *et al.* demonstrated that transcripts may also function by competing for microRNA binding, a biological activity independent of the translation of the protein they encode (Poliseno *et al.*, 2010). It is impossible for any currently available method, including ours, to assess which is the case.

The method described here is able to correctly classify as plausible a large fraction of the experimentally characterized isoforms, and to highlight dubious cases. Our aim is to provide easy access to a computational tool able to draw the attention of the life science community to them. Consequently, we took special care to convey the results of the analysis, although based on rather sophisticated tools, in an easy and understandable fashion. MAISTAS provides access to all the intermediate data used to generate the results, but it describes them in a human readable form. We believe that MAISTAS represents a step in the direction of using the knowledge accumulated in structural bioinformatics as well as the maturity of the tools available for applications related to the interpretation of genomic data and that it can be effectively used as a first step in characterizing novel proteins as well as a support for selecting interesting and intriguing cases for structural and functional studies.

ACKNOWLEDGEMENTS

We thank Loredana Le Pera, Andrea Sbardellati, Alejandro Giorgetti and Francesca Camilli for valuable feedback. We also thank Gianmauro Cuccuru, Michele Muggiri and Carlo Podda of the CRS4 High Performance Computing Group for their technical advice. We thank all the groups that kindly provided us with databases and binaries or source codes of the software installed and interfaced in this pipeline.

Funding: King Abdullah University of Science and Technology (KAUST; Award No. KUK-I1-012-43); Fondazione Roma and the Italian Ministry of Health (contract no.onc_ord 25/07, FIRB ITALBIONET and PROTEOMICA).

Conflict of Interest: none declared.

REFERENCES

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bairoch,A. *et al.* (2004) Swiss-Prot: juggling between evolution and stability. *Brief. Bioinformatics*, **5**, 39–55.

Battey,J.N. *et al.* (2007) Automated server predictions in CASP7. *Proteins*, **69** (Suppl. 8), 68–82.

Birney,E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Birzele,F. *et al.* (2008) ProSAS: a database for analyzing alternative splicing in the context of protein structures. *Nucleic Acids Res.*, **36**, D63–D68.

Cavallo,L. *et al.* (2003) POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level. *Nucleic Acids Res.*, **31**, 3364–3366.

Decker,W.K. and Craigen,W.J. (2000) The tissue-specific, alternatively spliced single ATG exon of the type 3 voltage-dependent anion channel gene does not create a truncated protein isoform *in vivo*. *Mol. Genet. Metab.*, **70**, 69–74.

Deutsch,E.W. *et al.* (2008) PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep.*, **9**, 429–434.

Durinck,S. *et al.* (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.

Fleming,P.J. and Richards,F.M. (2000) Protein packing: dependence on protein size, secondary structure and amino acid composition. *J. Mol. Biol.*, **299**, 487–498.

Flicek,P. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.

Hubbard,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.

Leinonen,R. *et al.* (2011) Improvements to services at the European Nucleotide Archive. *Nucleic Acids Res.*, **38**, D39–D45.

Leoni,G. *et al.* (2011) Coding potential of the products of alternative splicing in human. *Genome Biol.*, **12**, R9.

Maglott,D. *et al.* (2011) Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.

Martelli,P.L. *et al.* (2010) ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing. *Nucleic Acids Res.*, **39**, D80–D85.

Melamud,E. and Moult,J. (2009a) Stochastic noise in splicing machinery. *Nucleic Acids Res.*, **37**, 4873–4886.

Melamud,E. and Moult,J. (2009b) Structural implication of splicing stochastics. *Nucleic Acids Res.*, **37**, 4862–4872.

Pattabiraman,N. *et al.* (1995) Occluded molecular surface: analysis of protein packing. *J. Mol. Recognit.*, **8**, 334–344.

Poliseno,L. *et al.* (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, **465**, 1033–1038.

Pruitt,K.D. *et al.* (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Seal,R.L. *et al.* (2011) genenames.org: the HGNC resources in 2011. *Nucleic Acids Res.*, **39**, D514–D519.

Shionyu,M. *et al.* (2009) AS-ALPS: a database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse. *Nucleic Acids Res.*, **37**, D305–D309.

Soding,J. (2005) Protein homology detection by HMM–HMM comparison. *Bioinformatics*, **21**, 951–960.

The Uniprot Consortium (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.

Thierry-Mieg,D. and Thierry-Mieg,J. (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.*, **7** (Suppl. 1), s12: 1–14.

Tress,M.L. *et al.* (2007) The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl Acad. Sci. USA*, **104**, 5495–5500.

Voss,N.R. and Gerstein,M. (2005) Calculation of standard atomic volumes for RNA and comparison with proteins: RNA is packed more tightly. *J. Mol. Biol.*, **346**, 477–492.

Waterhouse,A.M. *et al.* (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.

Wilming,L.G. *et al.* (2008) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.*, **36**, D753–D760.

Paper III: A novel mechanism of natural vitamin E tocotrienol activity: involvement of Erβ signal transduction

Am J Physiol Endocrinol Metab 297: E427–E437, 2009.
First published June 2, 2009; doi:10.1152/ajpendo.00187.2009.

A novel mechanism of natural vitamin E tocotrienol activity: involvement of ERβ signal transduction

Raffaella Comitato,[1]* Kalanithi Nesaretnam,[2]* Guido Leoni,[1] Roberto Ambra,[1] Raffaella Canali,[1] Alessandro Bolli,[3] Maria Marino,[3] and Fabio Virgili[1]

[1]*National Research Institute for Food and Nutrition, Rome, Italy;* [2]*Malaysian Palm Oil Board, Selangor, Malaysia; and* [3]*Department of Biology, University Roma Tre, Rome, Italy*

Submitted 20 March 2009; accepted in final form 28 May 2009

Comitato R, Nesaretnam K, Leoni G, Ambra R, Canali R, Bolli A, Marino M, Virgili F. A novel mechanism of natural vitamin E tocotrienol activity: involvement of ERβ signal transduction. *Am J Physiol Endocrinol Metab* 297: E427–E437, 2009. First published June 2, 2009; doi:10.1152/ajpendo.00187.2009.—Vitamin E is a generic term used to indicate all tocopherol (TOC) and tocotrienol (TT) derivates. In the last few years, several papers have shown that a TT-rich fraction (TTRF) extracted from palm oil inhibits proliferation and induces apoptosis in a large number of cancer cells. However, the molecular mechanism(s) involved in TT action is still unclear. In the present study, we proposed for the first time a novel mechanism for TT activity that involves estrogen receptor (ER) signaling. In silico simulations and in vitro binding analyses indicated a high affinity of TTs for ERβ but not for ERα. In addition, in ERβ-containing MDA-MB-231 breast cancer cells, we demonstrated that TTs increase the ERβ translocation into the nucleus, which in turn activates estrogen-responsive genes (*MIC-1*, *EGR-1* and *cathepsin D*), as demonstrated by cell preincubation with the ER inhibitor ICI-182,780. Finally, we observed that TT treatment is associated with alteration of cell morphology, DNA fragmentation, and caspase-3 activation. Altogether, these experiments elucidated the molecular mechanism underling γ- and δ-TT effects.

estrogen receptor-β; breast cancer; apoptosis; tocopherol; nuclear receptor

TOCOTRIENOLS (TTs) are usually included together with tocopherols (TOCs) within the "vitamin E family". TTs have a chemical structure similar to TOCs but present three double bonds at positions 3′, 7′, and 11′ of the side chain. Similar to TOCs, TTs have four natural isomers, named as α, β, γ, and δ, that differ by the number and position of methyl groups on the chroman ring. The unsaturation of the side chain is associated with specific chemicophysical characteristics that are attracting growing interest both in the field of nutrition and in pharmacology (7, 62).

TOCs are commonly found in high concentrations in vegetable oils, animal fats, grains, vegetables, and fruits (13), whereas TTs are relatively rare in Western diets and found in appreciable levels only in a few specific vegetable fats, such as palm oil and rice bran oil (48). The great bioavailability of TOCs and their high efficiency in acting as antioxidants have attracted the interest of biologists who have disregarded TTs and their properties. Recent investigations have demonstrated that the antioxidant efficacy of TTs in membranes is higher than that of TOCs (59, 66), although their uptake and distribution after oral ingestion are less than that of α-TOCs (9, 76). Moreover, TTs have been reported to have many specific activities, such as the suppression of growth and the induction of apoptosis in different human and mouse mammary cancer cells (8, 20, 27, 35, 39–41, 57, 60, 67–69, 77) and in other human cancer cells (14, 15, 23, 37, 42, 63, 71). In general, TTs have been proposed to possess diverse properties that are often not exhibited by TOCs (58) and are associated with "anticancer" activity that is independent of their antioxidant properties.

The molecular mechanisms underlying these beneficial effects are still scarcely understood. In our laboratory, we have previously reported the effect of a TT-rich fraction (TTRF) from palm oil utilizing a cDNA gene array both on cultured human breast cancer cells (37) and in subcutaneously implanted athymic mice (38). This complex approach allowed the identification of a set of candidate genes involved in cell cycle control that are modulated at transcriptional level by TTRF. Therefore, we utilized these results as the background for further studies addressing the molecular mechanisms at the root of TT effects.

In this report, using a software-based approach confirmed by in vitro displacement experiments and by the use of the specific inhibitor for estrogen receptor (ER) ICI-182,780 in intact cells, we demonstrate that the effects of TTs are, at least in part, mediated by binding to the ERβ and involve the translocation into the nucleus of this nuclear receptor, inducing the expression of specific genes containing estrogen-responsive element (ERE) sequences in their promoter.

MATERIALS AND METHODS

Chemicals. TTRF was obtained from Sime Darby Plantation (Malaysia) and purified as described previously (65). Briefly, palm oil fatty acid distillate was converted into methyl esters by esterification. The methyl esters were then removed by distillation, leaving a vitamin E concentrate. This was further concentrated by crystallization and passed through an ion exchange column to give 60–70% pure vitamin E. Further purification was achieved by washing and then drying the concentrate, followed by a second molecular distillation stage. The final purity of vitamin E in TTRF was 95–99% and typically contained 32% α-TOC, 25% α-TT, 29% γ-TT, and 14% δ-TT.

Purified TTs were provided by Dr. Hiroyuki Yoshimura of Eisai Food and Chemical (Tokyo, Japan). Purity was ~99% for all TTs. Pure α-TOC was purchased from Sigma-Aldrich (St. Louis, MO). Stock solutions of TTRF and TTs were stored at −20°C in aliquots and diluted to the desired concentration in dimethyl sulfoxide (DMSO). The nonspecific ER antagonist ICI-182,780 was purchased from Tocris (Ballwin, MO).

Cell lines and treatments. MDA-MB-231 human breast cancer cells were obtained from American Tissue Culture Collection (Manassas, VA). Cells were grown in RPMI 1640 medium (Sigma-Aldrich)

*R. Comitato and K. Nesaretnam contributed equally to this article.

Address for reprint requests and other correspondence: R. Comitato, National Research Institute for Food and Nutrition, via Ardeatina 546, 00178 Rome, Italy (e-mail: comitato@inran.it).

supplemented with 10% fetal bovine serum (Sigma-Aldrich), penicillin/streptomycin (Invitrogen Life Science), 2 mM glutamine (Sigma-Aldrich), and 10% nonessential amino acid (MEM; Sigma-Aldrich).

Before any experimental session, cells were synchronized in G_1/G_0 by starvation in serum-free medium for 3 days. Once synchronized, 5.0×10^5 cells were seeded onto multiwell plates in phenol red-free RPMI 1640 and, where appropriate, incubated with ICI-182,780 (10^{-5} M in ethanol) for 30 min.

TTRF, purified TTs, or α-TOC were added to the culture medium for the indicated period of time. The actual concentration of TTs and α-TOC solutions was determined spectrophotometrically by using the extinction coefficients (ε_{292} α-TOC = 75.8, $\varepsilon_{292.5}$ α-TT = 91, ε_{296} γ-TT = 90.5, ε_{297} δ-TT = 89.1) before each experiment.

Final TTRF concentration in culture medium was set at 20 μg/ml to standardize the present study with our previous reports (41). Purified TTs and α-TOC concentrations are reported in Table 1. Concentrations of TTs used in the experiements were in the micromolar order, which is a concentration that can be achieved in the human serum, as reported previously (36, 53, 76). Control cells were treated with the same volumes of DMSO and/or ethanol vehicle alone.

After treatment, a Leitz Diavert microscope (Leitz, Weltzar, Germany) combined with a Nikon Coolpix P80 camera (Nikon, Tokyo, Japan) was used to check cell morphology and capture images.

Figures and tables present one of at least three independent experiments providing similar results or the mean (± SE) of at least three experiments, respectively.

Docking simulations. The Protein Data Bank identifiers of the three-dimensional structures used in this work are as follows: 3ERT (ERα antagonist), 1A52 (ERα agonist), 1NDE (ERβ antagonist), and 1U3Q (ERβ agonist) (downloadable from http://www.rcsb.org).

TTs (input for Autodock) were drawn with the Dundee PRODRG2 server (http://davapc1.bioch.dundee.ac.uk/programs/prodrg/). Crystallographic structures were "cleaned" by deleting water and solvent molecules, and hydrogen atoms were added with ADT tools (AutoDock). A minimization was made to relax the sterical contacts due to the introduction of hydrogen atoms using the "Minimize" module with "Amber99" force field included with the Thinker software. Finally, crystallographic ligands were deleted from any structure using ADT tools. As a test for AutoDock's ability to correctly map the chemical features of ER binding sites and the capacity to reasonably replicate known ligand-receptors interactions, biological ligand-receptor complexes from the selected crystallographic structures were separated and redocked. In all cases, AutoDock was found to be able to recover the known complex's structures and interactions.

Docking simulations were performed using AutoDock 4.0. For each receptor, the grid maps with the Autogrid 4.0 software using a grid box of 70 × 70 × 70 dimensions centered in the middle of the binding site were calculated. The docking simulations were performed with Lamarckian genetic algorithm requesting alternatively 20 and 90 GA runs with 250,000, 2,500,000, and 25,000,000 energy evaluations. The remaining parameters were kept to their default values. Cluster analysis was performed with the cluster function included in ADT tools.

Table 1. *Purity and concentration (μg/ml) of TTs used in experiments*

		Concentration (Times)	
	%TTRF	1	2
TTRF		20	
α-TT	25	5	10
γ-TT	29	5.8	11.6
δ-TT	14	2.8	5.6
α-TOC	32	6.4	12.8

TT, tocotrienol; TTRF, TT-rich fraction; TOC, tocopherol.

On the basis of the cluster analysis, we first performed an "intra-receptor conformation evaluation," observing for each docking simulation how the composition and the number of clusters obtained change upon the number of energy evaluations, and then a "between-receptors evaluation," searching the docking simulations that produce docking with few clusters, and are more populated with the low binding energy, and finally selecting the simulations where the most populated cluster also has the low binding energy.

TTRF and TT binding analyses. Values of the apparent molar fraction for 17β-estradiol (E_2), TTRF, and purified TTs binding to recombinant (PanVera, Madison, WI, USA) ERα or ERβ (Y^{app}) were determined by competitive radiometric binding assays using 4 nM [6,7-^{3}H(N)]estradiol (44.8 Ci/mmol, [^{3}H]-E_2; NEN Life Science, Boston, MA) as the tracer, as reported previously (11). Incubation was done at 25°C for 2 h in the binding buffer [0.04 M Tris·HCl, 1.5 mM EDTA, 1 mM DTT, 1% yeast extract (wt/vol), and 10% glycerol (vol/vol), pH 7.4]. The free and ERα- or ERβ-bound radioligands were separated by vacuum filtration through a 12-sample Millipore filter manifold (Millipore, Bedford, MA) holding glass microfiber filters (Whatman) (11). The radioactivity was counted with a 2100TR Tri-Carb liquid scintillation analyzer (Packard Instruments).

Values of the intrinsic molar fraction (Y) were obtained from Y^{app} values according to the following equation (11): $Y = Y^{app}/\{1 + ([B]/H)\}$, where [B] is the fixed [^{3}H]E_2 concentration (4 nM) and H is the equilibrium dissociation constant for [^{3}H]E_2 binding to ERα (0.2 ± 0.05 nM) or ERβ (0.27 ± 0.04 nM) at pH 7.4 and 25°C.

Values of the intrinsic equilibrium dissociation constant (K_d) for ligand binding to ERα or ERβ were obtained at pH 7.4 and 25°C according to a nonlinear four-parameter logistic model (28).

Nuclear localization of ERβ. In the Lab-Tek chamber slide system (Nalge Nunc International, Rochester, NY), 1×10^5 cells were treated with TTRF, α-TOC, and TTs for 24 h. At the end of experimental time, MDA-MB-231 cells were washed twice with PBS and fixed with paraformaldehyde buffer (3% paraformaldehyde in PBS with 0.5% Triton X-100) for 10 min at room temperature (RT). Cells were then washed three times with PBS and permeabilized with 100 μl of 0.2% Triton X-100 in PBS for 2 min at RT. Cells were then blocked with 4% BSA and 0.5% Triton X-100 in PBS for 30 min. Subsequently, cells were incubated with 20 μl of the diluted (1:20) ERβ primary antibody (Santa Cruz Biotechnology, Santa Cruz, CA) in blocking buffer for 2 h at RT. After four PBS washes, cells were incubated with diluted 1:20 secondary antibody goat anti-mouse fluorescein conjugated (Santa Cruz Biotechnology) for 1 h at RT. Cells were stained with 100 μl 4,6-diamidino-2-phenylindole to counterstain the nucleus. A Zeiss Axioskop II microscope (Carl Zeiss, Oberkochen, Germany) with appropriate filters was used. Images were collected and processed using the SPOT software.

RNA isolation. Total RNA was extracted from cells using TRI Reagent (Sigma-Aldrich) according to the manufacturer's instructions, with some minor modifications. Briefly, cells were homogenized in TRI Reagent solution and incubated for 5 min at RT. After the addition of 20% volume of chloroform, homogenates were vortexed for 2 min and centrifuged at 12,000 *g* for 15 min at 4°C. The resulting inorganic phase was collected. One volume of isopropanol at RT was added, and samples were stored on ice for 1 h. Finally, samples were centrifuged at 12,000 *g* for 30 min at 4°C, and the pellet was washed with 75% ethanol, centrifuged, dried, and resuspended in 50 μl of sterile water. RNA integrity was checked by denaturing gel electrophoresis, and concentration of each preparation was determined from $A_{260/280}$ reading using an undetermined ND-1000 spectrophotometer (NanoDrop, Wilmington, DE). Before real-time experiments, total RNA was treated with 25 units of DNase I (Clontech-BD Biosciences, Mountain View, CA) to remove any contaminating DNA.

Real-time PCR measurements. Primers (Table 2) corresponding to selected genes were designed with Primer Express 2.0 (Applied Biosystems, Foster City, CA). Real-time PCR was performed using

Table 2. *List of genes considered*

Gene	GenBank	Forward	Reverse
MIC-1	NM_004864	5′TGGTGCTCATTCAAAAGACCG3′	5′GTGGAAGGACCAGGACTGCTC3′
EGR-1	NM_001964	5′CTCCACAGGGCTTTGGGAC3′	5′GAGAGGGAGGACTTGGCTCTG3′
Cathepsin D	NM_001909	5′CTGTGAGGCCATTGTGGACAC3′	5′CAGCTTGTAGCCTTTGCCTCC3′
Bmp-4	NM_001202	5′GCCGTCATTCCGGACTACAT3′	5′GGGCGCTCAGGATACTCAAG3′
β-Actin	NM_001614	5′AGAAGGATTCCTATGTGGGGG3′	5′CATGTCGTCCCAGTTGGTGAC3′

MIC-1, macrophage inhibitory cytokine-1; EGR-1, early growth response factor-1; Bmp-4, bone morphogenetic protein-4. GenBank identification code and sequence of primers utilized for real-time PCR.

the SuperScript Platinum SYBR Green One-Step kit (Invitrogen). Quantitative PCR was conducted in duplicate in a 15-μl reaction volume containing 7.5 μl of 2× SYBR mix, 0.3 μl of SuperScript Platinum III RT *Taq* mix, 0.3 μl of primer F (10 μM), 0.3 μl of primer R (10 uM), and sterile water. Finally, 5 μl of opportunely diluted RNA was added. Reactions were carried out in an ABI Prism 7900HT (Applied Biosystems). Thermal protocol was 65°C for 1 min, 50°C at 20 min, and 95°C at 5 min, followed by 40 cycles of two steps of denaturation at 95°C for 15 s and extension at 60°C for 30 s. Fluorescence data was collected at the extension step and given as threshold cycle (C_T). The C_T values for each target and reference genes were obtained, and their difference was calculated (ΔC_T). For normalization purpose, an identical set of reactions was prepared using primer specific for β-actin. Quantitative differences in the cDNA target among samples were determined using the mathematical model of Pfaffl (52), in which an expression ratio was determined for each sample by calculating $(E_{target})^{\Delta C_T(target)}/(E_{\beta\text{-actin}})^{\Delta C_T(\beta\text{-actin})}$, where E is the efficiency of the primer set and $\Delta C_T = C_T$ (β-actin) − C_T (target). Finally, the results were log2 transformed to obtain data distributed symmetrically. The amplification efficiency of each primer set was calculated from the slope of a standard amplification curve of log microliter cDNA/reaction vs. C_T value over at least four orders of magnitude [$E = 10^{-(1/\text{slope})}$]: β-actin primers, $E = 2.15$; early growth response factor-1 (EGR-1) primers, $E = 1.90$; Macrophage inhibitory cytokine-1 (MIC-1) primers, $E = 2.03$; cathepsin D primers, $E = 1.86$; bone morphogenetic protein-4 (Bmp-4) primers, E = 1.97.

Protein extraction and Western blot. Cells were lysed in RIPA buffer containing 50 mM Tris·HCl (pH 7.5), 150 mM NaCl, 1% NP-40, 0.5% sodium deoxycholate, 2 mM sodium fluoride, 2 mM EDTA, 0.1% SDS, and an EDTA-free protease inhibitor cocktail (Sigma-Aldrich) for 1 h on ice. Cell extracts were centrifuged at 14.000 *g* for 15 min at 4°C, and the supernatants were used for Western blot experiments. Protein concentrations were determined with a protein assay kit (Bio-Rad Laboratories, Hercules, CA) at 596 nm wavelength. Protein samples (30 μg/lane) were mixed with 6× SDS reducing sample buffer and boiled for 5 min before loading. Proteins were separated in SDS polyacrylamide gels and transferred to polyvinylidene difluoride (PVDF) membranes (Millipore). Membranes were blocked with 5% nonfat milk dried (AppliChem, Darmstadt, Germany) in PBS-T (1% PBS and 0.1% Tween-20) for 1 h at room temperature and then incubated overnight at 4°C with a 1:500 dilution of rabbit EGR-1 antibody (Santa Cruz Biotechnology), 1:500 cathepsin D rabbit antibody (Santa Cruz Biotechnology), 1:250 MIC-1 goat antibody (Novus Biologicals, Littleton, CO), 1:500 Bmp-4 (Santa Cruz Biotechnology), 1:2,000 caspase 3 goat antibody (Santa Cruz Biotechnology), 1:400 ERα (Santa Cruz Biotechnology), 1:1,000 ERβ (Santa Cruz Biotechnology), and 1:1,000 α-tubulin mouse antibody (MP Biomedicals, Irvine, CA). After three washings with PBS-T, membranes were incubated for 1 h at RT with 1:2,000 goat anti-mouse or goat anti-rabbit or 1:5,000 donkey anti-goat peroxidase-conjugated secondary antibodies (Santa Cruz Biotechnology). After three washings with PBS-T, specific spots were detected by chemiluminescence reagent ECL Plus (Amersham Pharmacia Biotech, Piscataway, NJ) and visualized by autoradiography with high-performance chemiluminescence film (Amersham Biosciences, Buckinghamshire, UK).

DNA laddering. Protocol was adapted from Gooch and Yee (19). After 24 h, adherent and nonadherent cells were collected, centrifuged at 12,000 *g* for 5 min, washed with 1× PBS, pelleted again, and then lysed in cold 0.15 M NaCl, 10 mM Tris·HCl (pH 7.8), 2 mM $MgCl_2$, 1 mM DTT, and 0.5% NP-40 on ice for 1 h. Lysates were centrifuged at 12,000 *g* for 10 min, and pellets (nuclei) were resuspended in cold 0.35 M NaCl, 10 mM Tris·HCl (pH 7.8), 1 mM $MgCl_2$, and 1 mM DTT and stored at −80°C overnight. Lysates were then extracted once with phenol-chloroform, and DNA was precipitated with 0.01 M $MgCl_2$ and 2.5 volumes of 100% ethanol overnight at −20°C. DNA was collected by centrifugation at 16,000 *g* for 30 min, resuspended in 10 mM Tris·HCl and 1 mM EDTA plus 0.1 mg/ml RNase A, and incubated at 37°C for 1 h. Proteinase K (1 mg/ml) was added for an additional hour at 37°C. DNA was then electrophoresed in 1.5% agarose gels containing ethidium bromide and visualized by UVIpro Bronze acquisition system (UVITEC, Cambridge, UK).

Statistics. All data represent means ± SE of at least three independent experiments. Statistical analysis was performed with R software from the R Foundation for Statistical Computing (Vienna, Austria). Real-time data were analyzed by one-way ANOVA with repeated measures followed by the Bartlett and Fligner-Killeen test for homogeneity of variance. Dunnett post hoc test was used to evaluate differences among multiple conditions. Binding data were analyzed with Student's *t*-test. *P* values ≤0.05 were considered to be statistically significant.

RESULTS

A deeper analysis of results obtained from previous cDNA arrays (see Refs. 37 and 38), together with new ones addressing the effect of a stripped culture medium characterized by total removal of lipid components of serum, including estrogens (data not shown), indicated the possibility of reconsidering the hypothesis that the effect of TTRF on gene expression could result from interaction with ER-related signaling.

Therefore, the experiments hereby presented have been designed to either confirm or discard the novel hypothesis of the involvement of ERs in TTRF and TT's effects on gene expression.

Docking simulations and ligand binding to ERs. We first studied the interaction between TTs and ERs by means of a software-based approach, performing docking simulations in two distinct states, agonist and antagonist bound. Simulations done with the AutoDock 4.0 software (see MATERIALS AND METHODS) strongly suggest a high likelihood of the binding of δ-TT and γ-TT rather than α-TT to ERβ (Fig. 1). This difference in TT activity follows the number of substitutions in their phenolic ring, and it is a common feature of other estrogen and estrogen-like molecules (4). On the other hand, in silico simulations also suggest a low affinity for TT binding to

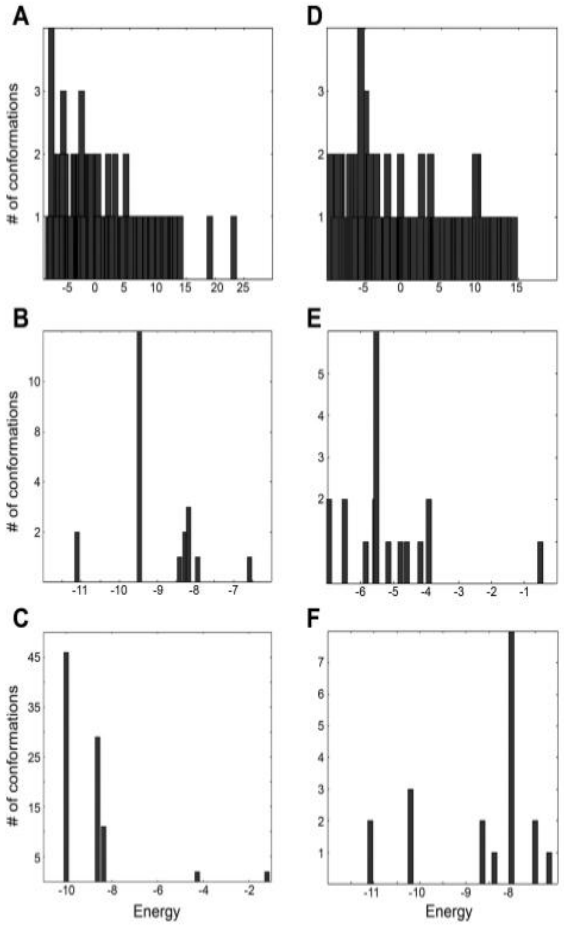
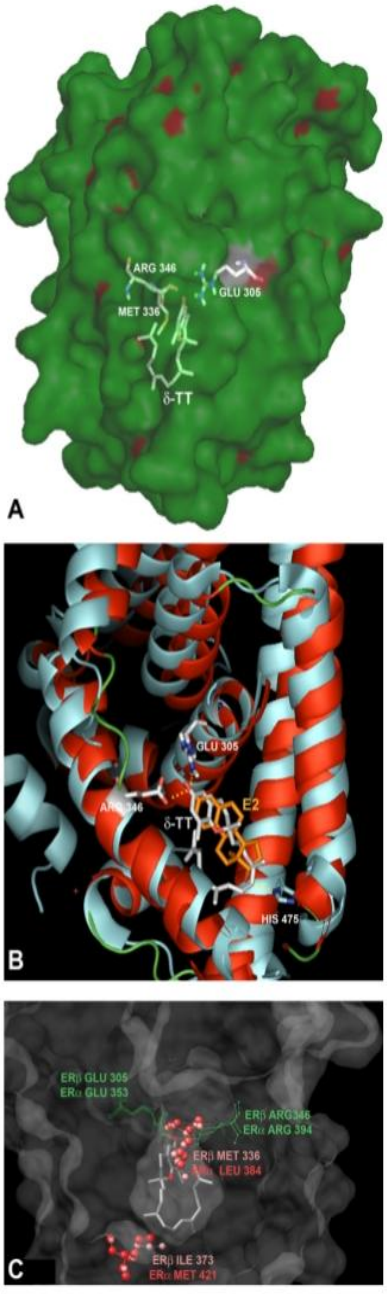


Fig. 1. Best clustering results (25,000,000 E evaluations) are shown. All results shown assume the receptor in the agonist-bound conformation, and the number of requesting solutions is as follows: α-tocotrienol (TT)-estrogen receptor (ER)β, 90 (*A*); γ-TT-ERβ, 20 (*B*); δ-TT-ERβ, 90 (*C*); α-TT-ERα, 90 (*D*); γ-TT-ERα, 20 (*E*); δ-TT-ERα, 20 (*F*). The minor number of clusters and the lowest energy of the most populated cluster were the criteria for the individuation of the "best docking." Overall, our results suggest that γ-TT and δ-TT, rather than α-TT, bind better with a global higher affinity for ERβ compared with ERα.

ERα. The strength of interaction was compared with the in silico reconstructed E_2-ERβ complex. Our docking data indicate a molecular mechanism involving a hydrogen bond between the phenolic oxygen of TT and Arg^{346} and Glu^{305} of ERβ ligand-binding domain (Fig. 2*A*). This bond is analogous to the characteristic hydrogen bond between estrogens and the Arg-Glu-H_2O triad in the biological (endogenous or classical antiestrogen molecules) ligand-receptor complexes (Fig. 2*B*). Moreover, our results show that the TT semiplanar/aromatic surface is orientated in a favorable position to establish Van der Waals interaction with Met^{336} (which is substituted in ERα by Leu^{384}, as indicated above). This feature appears to be common to other planar/aromatic ERβ ligands, including some phytoestrogens such as genistein or other synthetic molecules such as Benzopyrans (33, 45). The Van der Waals interaction has to be added to another beneficial interaction that is represented by the "hydrophobic effect" due to the presence of Ile^{373} in ERβ (instead of Met^{421} in ERα), which could better accept the tocolic (isoprenoid) tail of TTs (Fig. 2*C*).

These data have been confirmed by in vitro binding displacement test analyses. Data reported in Fig. 3 (○) indicate that K_d value for E_2 binding to recombinant ERα is 0.21 ± 0.02 nM (Fig. 3*A*), whereas neither TTRF as a whole nor α-TT, δ-TT, or γ-TT bind to this receptor isoform (Fig. 3, *B*, *C*, *D*, and *E*, respectively). According to thermodynamic considerations (11, 28), the K_d value for TTRF, TTs, and α-TOC binding to ERα is >100 μM. On the other hand, K_d values obtained by using recombinant ERβ (Fig. 3, ●) designate TTRF, γ-TT, and δ-TT as specific ligands for this receptor isoform. In fact, K_d value for E_2 is 0.27 ± 0.05 nM (Fig. 3*A*′), for TTRF is 0.65 ± 0.05



μM (Fig. 3*B'*), for δ-TT is 0.110 ± 0.012 μM (Fig. 3*D'*), for γ-TT is 0.130 ± 0.015 μM (Fig. 3*E'*), and for α-TOC is 78 ± 5.3 nM, whereas the K_d value for α-TT binding to ERβ is >100 μM (Fig. 3*C'*), confirming and corroborating docking simulation data (Fig. 1).

Altogether, these data indicate that TTRF, δ-TT, and γ-TT associate with ERβ, although do not provide any information about the ability of these compounds to modulate ERβ activities. However, as indicated by the intrareceptor evaluation (see MATERIALS AND METHODS), the preferential receptor conformation is the one having the helix 12 in the agonist-bound conformation, suggesting an activating effect of TTs on ERβ activities.

Effects of TTRF and TT on ERβ transcriptional activity. To corroborate the agonist interaction between TTs and ERβ suggested by docking simulation, we analyzed the cellular localization of this receptor isoform following TTRF, α-TOC, and TT treatment in MDA-MB-231 breast cancer cells by indirect immunofluorescence. MDA-MB-231 cell line expresses both ERα and ERβ transcripts (data not shown) (70), but only ERβ is expressed at the level of protein (Fig. 4). As shown in fig. 5, ERβ localization in control MDA-MB-231 cells was predominantly cytoplasmatic, according to previous reports (25, 75), with nuclei only slightly stained. Treatment with TTRF was associated with nuclear staining, and the administration of purified γ-TT or δ-TT induced a more marked signal than the mixture (Fig. 5). On the other hand, no effects were observed in cells treated with α-TT, in agreement with docking simulation and in vitro ligand binding to the purified protein assays. Finally, no effects were observed in cells treated with α-TOC, in agreement with in silico simulations.

Thus we tested the expression of the ERE-containing genes *MIC-1*, *EGR-1*, and *cathepsin D* in MDA-MB-231 cells at the baseline, after treatment with different concentrations of TTRF alone, or in the presence of ERβ-specific inhibitor (ICI-182,780) pretreatment. This set of genes was taken from our database obtained from previously reported gene array-based studies (37). TTRF treatment induced a significant increase of the expression of *MIC-1*, *EGR-1*, and *cathepsin D* mRNAs at both 24 and 48 h (Fig. 6, *A* and *B*, respectively). Cell pretreatment with ICI-182,780 partially or totally prevented the upregulation of all analyzed gene expression. Moreover, as shown in Fig. 7, *MIC-1* and *EGR-1* gene expression increased after γ-TT and δ-TT administration only at doses corresponding to two times the amount of TTs present in 20 μg/ml of TTRF. The pretreatment with ICI-182,780 decreased the upregulation of the expression of all genes at all of the concentrations considered (data not shown), demonstrating the involvement of ERβ in TT's transcriptional effects. As expected, the expression of an ERE-devoid gene, *Bmp-4*, was not af-

Fig. 2. *A*: 3-dimensional model of the lowest energy complex between δ-TT and ERβ in agonist-bound conformation. *B*: superposition between crystallographic structures of ERβ (RCSB protein databank code: 1X7J) bound to estradiol is shown in cyan and our best docking complex between ERβ and δ-TT in red. *C*: superposition of ERβ-ERα binding sites. Most relevant residues are colored according to their hydropatic properties (in green are the most hydrophilic amino acids; in red are the most hydrophobic amino acids). The hydropathy index is calculated on the basis of the Kyte-Doolittle scale by the Chimera software. E_2, 17β-estradiol.

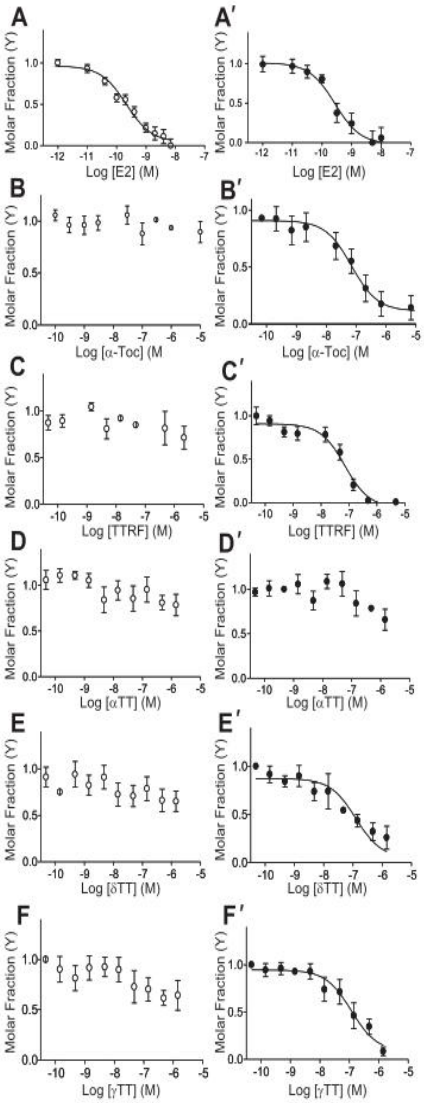


Fig. 3. Competitive radiometric binding of free E_2 (*A* and *A′*), α-tocopherol (α-TOC; *B* and *B′*), TT-rich fraction (TTRF; *C* and *C′*), α-TT (*D* and *D′*), δ-TT (*E* and *E′*), and γ-TT (*F* and *F′*) to ERα (○) and ERβ (●) showing the dependence of the intrinsic molar fraction (*Y*) on ligand concentrations. The continuous lines were obtained as described by Kuiper et al. (28). For details, see text.

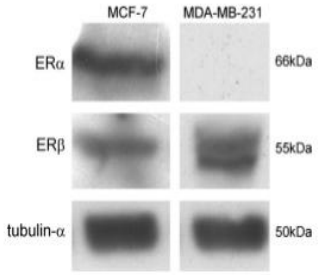


Fig. 4. Cellular proteins were isolated from MDA-MB-231 run in acrylamide gels; the levels of ERα and ERβ proteins were assessed by Western blot, and immunoreactivity was revealed using ECL Plus (Amersham). As shown, no ERα protein could be detected in MDA-MB-231 cells. One of 3 independent experiments yielding comparable results is shown.

fected by TTs at any of the experimental conditions. Finally, neither α-TOC nor α-TT treatments were able to significantly affect the expression of *MIC-1* and *EGR-1* genes, even at the highest doses.

Protein levels were consistent with mRNA levels. Cathepsin D, MIC-1, and EGR-1 levels increased significantly by TTRF (Fig. 8) and TT (data not shown) treatment and remained stable at the baseline level in the presence of ICI-182,780 pretreatment (Fig. 8). As expected, protein levels of the ERE-devoid gene *Bmp-4* were not affected by treatments.

Our data indicate that the effects of specific TTs on gene expression in MDA-MB-231 cells are due, at least in part, to their interaction on the ERβ transcriptional pathway.

Effects of TTRF and TTs on apoptosis. Finally, in our studies, we observed that treatment of MDA-MB-231 cells with TTs was associated with alteration of cell morphology (Fig. 9). In fact, the MDA-MB-231 cells lost their spindle-shaped morphology and became smaller and rounded. We also observed detached, flattened, and swollen cells. Trypan blue staining indicated that the majorities of floating cells were not necrotic and still alive (data not shown), suggesting that cells might start apoptosis. To determine whether TTRF and TTs may induce apoptosis, we extracted DNA from adherent and nonadherent cells and then subjected it to agarose gel electrophoresis. Figure 10 shows that DNA fragmentation occurs in MDA-MB-231 cells treated with either TTRF or different concentrations of purified TTs. In particular, we observed the typical apoptotic DNA ladders at the concentration of γ- and δ-TTs (γ-TT^2 and δ-TT^2) equal to two times the amount provided by 20 μg/ml TTRF. This value is in the range of 10–20 μM, which is to be considered a high, albeit still physiologically achievable, concentration. These data have been confirmed by the Western blot analysis of caspase-3 proform (32-kDa band) cleavage that results in the production of the active subunit of the protease (17-kDa band). As shown in Fig. 10, γ-TT, δ-TT, and TTRF treatments were associated with caspase-3 activation, one of the key factors involved in mitochondrial-dependent apoptosis. On the other hand, the treatment with either α-TOC or α-TT, which did not bind to ERβ, was not associated with DNA laddering, suggesting that ERβ is also involved in this TTRF effect.

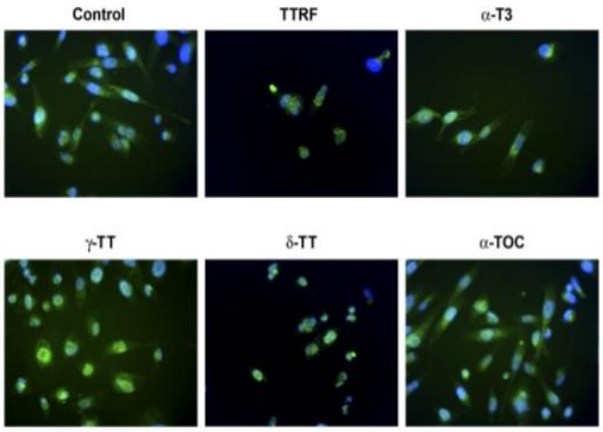


Fig. 5. Immunofluorescence of MDA-MB-231 cells using ERβ antibody. Cells were treated for 24 h with TTRF (20 μg/ml), purified TTs, or α-TOC (at the 1× concentration as in Table 2). TT treatment was associated with strong nuclear staining for TTRF and especially for purified γ-TT or δ-TT. On the other hand, little or no nuclear internalization was observed in cells treated with α-TT or α-TOC. One of 3 independent experiments yielding comparable results is shown. MIC-1, macrophage inhibitory cytokine-1; ICI, ICI-182,780.

DISCUSSION

TTs have recently gained increasing scientific interest due to their specific activity pattern. Available studies suggest that TTs and TOCs are metabolized similarly, although TTs have been reported to be degraded to a larger extent than TOCs (9, 76). In humans, the plasma levels of all TTs increase markedly up to μmol/l levels following the intake of TT-rich food and then drop rapidly, followed by a more gradual decline. TTs have been found in various tissues of rats, especially adipose tissues, skin, and heart, after oral application, suggesting that these molecules are absorbed and distributed in vivo (56).

Several reports have indicated that both TOCs and TTs have an antitumoral activity independent of their antioxidant activity (10, 12, 24, 49). Therefore, several authors have focused their interest on the molecular pathways involved in cellular response induced by TOCs and TTs and reported a modulation of different intracellular signaling pathways (1, 3, 27, 35, 54, 57, 61). In the present study, a direct interaction of a TT-rich extract from palm oil (TTRF) and purified TTs with the ERβ has been demonstrated by combining software-based docking simulations, in vitro E_2 displacement assays, and immunocytochemistry and by evaluating ERE-dependent gene expression in MDA-MB-231 cells, a line of human breast cancer cells expressing ERβ but not ERα.

We have reported previously that TTRF inhibits the growth of breast cancer cells expressing different ER profiles both cultured in vitro and after implantation in athymic nude mice (40, 41). The evidence that TTRF activity was not associated with a specific ER expression profile led us to the initial conclusion that TTRF activity was independent of estrogen-related signaling. Similar conclusions have been reported by Guthrie et al. (20).

Our data suggest that the conclusions drawn from such studies, based on the assumption of a total lack of ER-dependent signaling in the MDA-MB-231 cell line, are to be reconsidered. This cell line, obtained from an advanced human breast adenocarcinoma, is in fact often believed to be void of any ER activities, and its utilization is considered an expedient experimental model to study the modulation of cell proliferation independently of ER-mediated mechanisms. On the contrary, these cells have frequently been found to express both ER transcripts and only high levels of ERβ protein (26, 30, 31, 70).

Software simulations indicate that ERβ could interact with different TT isoforms with different intensities through a hydrogen bond between the phenolic hydroxyl group and the Arg^{346}-Glu^{305} residues in the binding cavity according to the well-known interaction between the triad Arg-Glu-H_2O and

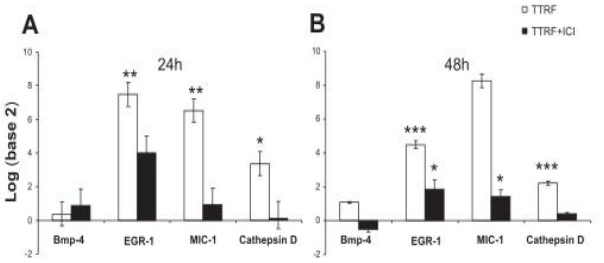


Fig. 6. Expression of bone morphogenetic protein-4 (Bmp-4), early growth response factor-1 (EGR-1), MIC-1, and cathepsin D genes in MDA-MB-231 cells treated for 24 and 48 h with 20 μg/ml TTRF after ICI-182,780 pretreatment for 30 min. Gene expression was analyzed by real-time quantitative PCR, and results were log transformed (logarithm 2) to obtain data distributed symmetrically. Statistical significance was calculated by Dunnett post hoc test (***$P \leq 0.001$; **$P \leq 0.01$; *$P \leq 0.05$). Data showed represent the pooled values of 3 independent experiments.

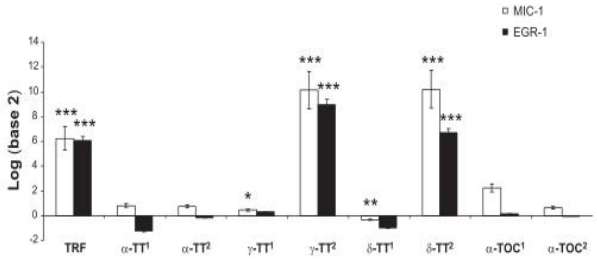


Fig. 7. Comparison of TTRF and single TTs or α-TOC effects in MDA-MB-231 cells (concentrations are intended 1 or 2 times, as in Table 2) in terms of expression of MIC-1 and EGR-1 genes. Gene expression was analyzed by real-time quantitative PCR, and results were log transformed (logarithm 2) to obtain data distributed symmetrically. Statistical significance was calculated by Dunnett post hoc test (***$P \leq 0.001$; **$P \leq 0.01$; *$P \leq 0.05$). Data showed represent the pooled values of 3 independent experiments.

phenolic group of ER antagonists or agonists. This evidence has been confirmed by competitive binding assays, which indicate that TTRF and TTs possess a higher binding affinity for ERβ with respect to ERα. These differences, as suggested by the in silico simulations, apparently rely on the presence of a Met in position 336 (which is substituted in ERα by Leu[384]). This residue is able to generate Van der Waals interactions with the semiplanar/aromatic surface of TTs. This feature is actually shared with other planar/aromatic ERβ ligands, including some phytoestrogens as genistein or other synthetic molecules, including Benzopyrans (33, 45). Overall, in silico docking studies and the K_d data are consistent in indicating a higher binding affinity between ERβ and δ-TT and γ-TT compared with α-TT, possibly reflecting the different dimensions of ligands due to methyl groups on their chromanolic ring and, therefore, a larger steric hindrance.

Ligand-activated ERs dimerize and translocate in the nucleus, where they recognize specific EREs located in or near promoter DNA regions of target genes (2, 22, 44). Behind this direct genomic mechanism, shared with other steroid hormones, E_2 also modulates gene expression by a second indirect mechanism that involves the interaction of ERs with other transcription factors, which in turn bind their cognate DNA elements. In this case, ER modulates the activities of transcription factors such as the activator protein-1 (AP-1) (18) or stimulating protein-1 (46) by stabilizing DNA-protein complexes and/or recruiting coactivators.

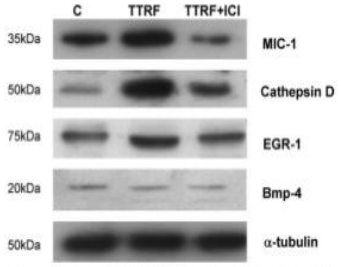


Fig. 8. Total proteins from MDA-MB-231 cells treated for 24 h with 20 μg/ml TTRF after ICI-182,780 pretreatment for 30 min were separated in 10% SDS polyacrylamide gels and transferred to PVDF membranes (Amersham). Ponceau-S staining was used to assess that equivalent amounts of protein were loaded on each lane (not shown). The blots were probed with specific antibodies, and immunoreactivity was revealed using ECL Plus (Amersham). One representative out of 3 experiments performed on is shown. C, control.

Immunocytochemistry experiments performed on MDA-MB-231 cells indicate that both TTRF and TT treatment are associated with ERβ translocation into the nucleus, in logical agreement with both in silico simulations and in vitro displacement experiments. In fact, both γ-TT and δ-TTs treatment were associated with ERβ nuclear transfer, whereas neither the treatment with α-TT nor with α-TOC induced ERβ translocation. Amazingly, despite the evidence of a high-affinity binding of α-TOC on the purified recombinant ERβ, we could not observe nuclear translocation of ERβ nor other cellular responses. Wang et al. (73) recently reported that 4-hydroxytamoxifen (HT) can occupy not only the core binding pocket within the ligand-binding domain of ERβ but also a second site that overlaps the hydrophobic groove of the coactivator recognition surface. This observation suggests that, similarly to HT, other small molecules such as α-TOC can act as direct antagonists of receptor-coactivator interaction, therefore impairing some of the ERβ-mediated cellular response. This possibility is under active consideration in our laboratories.

Both TTRF and purified γ-TT and δ-TT induce only the expression of genes bearing ERE sequences in their promoters, namely *MIC-1*, *EGR-1*, and *cathepsin D* in MDA-MB-231 cells. This positive regulation is countered by ICI-182,780, indicating that ERβ is a specific mediator of TT effects. ICI-182,780 molecule is in fact a well-known specific estrogen antagonist blocking the majority of pathways mediated by ERα and ERβ (32, 51). Notably, α-TOC and α-TT did not induce any significant changes in gene expression of *MIC-1* and *EGR-1*. It is interesting to note that MDA-MB-231 cells have a constitutive high level of activation of the transcription factor AP-1 (data not shown), which was not increased further following treatment with either TTRF or purified TTs. This observation strengthens the indication of a direct interaction of activated ERβ and EREs rather than an indirect interaction with DNA through jun-fos proteins (5, 50). These data are in agreement with the docking and immunocytochemistry analysis demonstrating that the components of TTRF responsible for ERβ activation are γ- and δ-TT. This specificity is in agreement with previously published studies by other authors that showed a more evident inhibitory effect on cell growth by TTs (39, 41) and a marked proapoptotic effect of TTs (3, 55, 61).

The contribution of ERβ to E_2-induced cell proliferation is very different to that elicited by ERα. ERβ appears to act as a negative dominant regulator in E_2 signaling, and when coexpressed with ERα it causes a concentration-dependent reduction of ERα-mediated transcriptional activation and the repression of ERα-mediated effects, including cell proliferation (47).

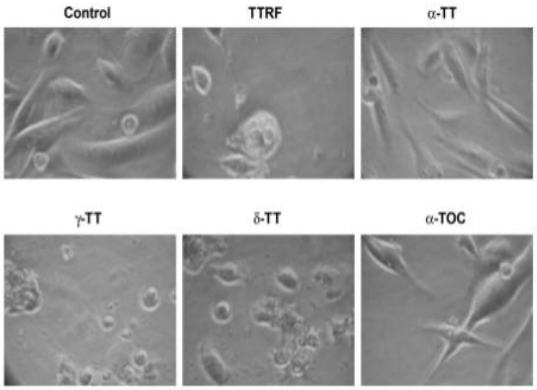


Fig. 9. TT treatment was associated with alteration of cell morphology. MDA-MB-231 cells treated for 24 h with TTRF (20 μg/ml), purified TTs, or α-TOC (at the 1× concentration, as in Table 2) were analyzed using a Leitz Diavert microscope and captured with a Nikon Coolpix P80 camera. TTRF, γ-TT, and δ-TT induced loosening of the characteristic spindle shape and detaching and flattening of cells. On the other hand, no effect was observed in cells treated with α-TT or α-TOC. One of 3 independent experiments yielding comparable results is shown.

Consistent with this notion, E_2 increases cell proliferation and causes tumor formation in MCF-7 cells expressing higher levels of ERα compared with ERβ (34). On the other hand, the presence of transfected ERβ is associated with the inhibition of E_2-induced proliferation in MCF-7 cells, and, in the presence of ERβ, E_2 rapidly induces p38 MAPK activation, which in turn leads to caspase-3 activation and poly(ADP-ribose) polymerase cleavage in colon cancer cells (6, 17). Therefore, it is highly probable that selective ERβ ligands drive MDA-MB-231 breast cancer cells through apoptosis. According to this hypothesis, we observed that both γ- and δ-TT induced morphological changes of cells associated with cell death, DNA laddering, and caspase-3 activation. This observation matches with the observed *MIC-1* upregulation by TTRF, γ-TT, and δ-TT treatments. In fact, it has previously been reported that *MIC-1*, which is a divergent member of the "transforming growth factor-β" superfamily, induces caspase-dependent apoptosis and reduces cell matrix and cell-cell adhesion in hormone-sensitive prostate cancer cells (29). Moreover, our preliminary data indicate a delay in the S/G_2 transition in breast cancer cells treated with TTRF (data not shown). These results are in agreement with data reported by Agarwal et al. (2), who observed a significant MIC-1 overexpression leading to S-phase arrest and apoptosis in ovarian cancer cells treated with *N*-(phosphonacetyl)-L-aspartate, a potent and reversible inhibitor of pyrimidine nucleotide synthesis in cells. Thus, we hypothesize that apoptosis induced by γ- and δ-TT can be regulated, at least in part, by the ERβ signaling pathway and increasing *MIC-1* level, eventually leading to a reduction of cell adhesion, followed by cell cycle arrest and, finally, caspase-dependent apoptosis.

The hypothesis of an interaction between specific forms of different molecules grouped under the generic term "vitamin E" and nuclear receptors is actually supported by other indications. For example, several functions affected by vitamin E status in experimental animals such as those involved in the reproductive performances and fetus reabsorption are obviously controlled by ER-dependent responses (16, 72, 74). Moreover, Ni et al. (43) and Zhang et al. (78) demonstrated that TOCs are able to modulate prostate cancer cell growth, and this involves inhibition of androgen receptor expression (78). Finally, recent reports indicate that γ-TT suppresses MAPK (64) and phosphatidylinositol 3-kinase/phosphoinositide-dependent protein kinase-1/Akt signaling (61), which act as second messengers in ERα-dependent "nongenomic" activity of steroid hormone.

In conclusion, even if other factors and alternative pathways could be involved in TT signaling, we have strong original

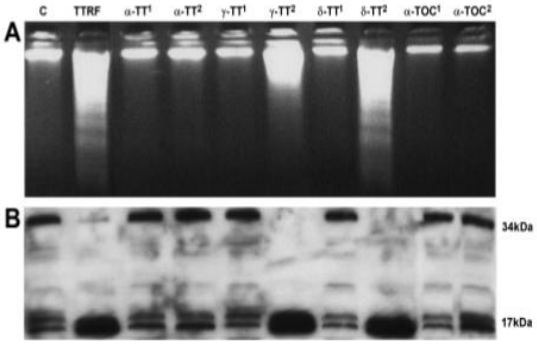


Fig. 10. TTs induce apoptosis in MDA-MB-231 cells. Cells were treated for 24 h with TTRF or purified TTs or α-TOC (concentration is intended 1 or 2 times, as in Table 2), and then DNA and proteins were extracted. *A*: DNA was electrophoresed in 1.5% agarose gels containing ethidium bromide and visualized by UVIpro Bronze acquisition system (UVITEC, Cambridge, UK). The typical laddering was observed for TTRF γ-TT and δ-TT only. *B*: total proteins were separated in 10% SDS polyacrylamide gels and transferred to PVDF membranes (Amersham). Ponceau-S staining was used to assess that equivalent amounts of protein were loaded on each lane (not shown). The blots were probed with the specific caspase-3 goat antibody (Santa Cruz Biotechnology), and immunoreactivity was revealed using ECL Plus (Amersham). One representative out of 3 experiments performed is shown.

indications supporting a role of TTs as efficient and selective ligands of ERβ. In our experimental model, the interaction between γ- and δ-TT with ERβ is productive and modulates its transcriptional activity through ERE. Our observations open new avenues for a specific role for vitamin E forms in regulating gene expression and in modulating cancer cell growth.

ACKNOWLEDGMENTS

We thank Dr. Hiroyuki Yoshimura of Eisai Food and Chemical for the generous gifts of purified TT.

GRANTS

This work was supported by the Malaysian Palm Oil Board, the National Institute of Biostructures and Biosystems, and the Ministry of Education, University, and Research of Italy (FISR "Safe-Eat").

REFERENCES

1. **Agarwal MK, Agarwal ML, Athar M, Gupta S.** Tocotrienol-rich fraction of palm oil activates p53, modulates Bax/Bcl2 ratio and induces apoptosis independent of cell cycle association. *Cell Cycle* 3: 205–211, 2004.
2. **Agarwal MK, Hastak K, Jackson MW, Breit SN, Stark GR, Agarwal ML.** Macrophage inhibitory cytokine 1 mediates a p53-dependent protective arrest in S phase in response to starvation for DNA precursors. *Proc Natl Acad Sci USA* 103: 16278–16283, 2006.
3. **Ahn KS, Sethi G, Krishnan K, Aggarwal BB.** Gamma-tocotrienol inhibits nuclear factor-kappaB signaling pathway through inhibition of receptor-interacting protein and TAK1 leading to suppression of antiapoptotic gene products and potentiation of apoptosis. *J Biol Chem* 282: 809–820, 2007.
4. **Anstead GM, Carlson KE, Katzenellenbogen JA.** The estradiol pharmacophore: ligand structure-estrogen receptor binding affinity relationships and a model for the receptor binding site. *Steroids* 62: 268–303, 1997.
5. **Aranda A, Pascual A.** Nuclear hormone receptors and gene expression. *Physiol Rev* 81: 1269–1304, 2001.
6. **Ascenzi P, Bocedi A, Marino M.** Structure-function relationship of estrogen receptor alpha and beta: impact on human health. *Mol Aspects Med* 27: 299–402, 2006.
7. **Atkinson J, Epand RF, Epand RM.** Tocopherols and tocotrienols in membranes: a critical review. *Free Radic Biol Med* 44: 739–764, 2008.
8. **Azzi A, Breyer I, Feher M, Ricciarelli R, Stocker A, Zimmer S, Zingg J.** Nonantioxidant functions of alpha-tocopherol in smooth muscle cells. *J Nutr* 131: 378S–381S, 2001.
9. **Birringer M, EyTina JH, Salvatore BA, Neuzil J.** Vitamin E analogues as inducers of apoptosis: structure-function relation. *Br J Cancer* 88: 1948–1955, 2003.
10. **Blot WJ, Li JY, Taylor PR, Guo W, Dawsey S, Wang GQ, Yang CS, Zheng SF, Gail M, Li GY, Yu Y, Liu BQ, Tangrea J, Sun YH, Liu F, Fraumeni JF, Zhang YH Jr, Li B.** Nutrition intervention trials in Linxian, China: supplementation with specific vitamin/mineral combinations, cancer incidence, and disease-specific mortality in the general population. *J Natl Cancer Inst* 85: 1483–1492, 1993.
11. **Bolli A, Galluzzo P, Ascenzi P, Del Pozzo G, Manco I, Vietri MT, Mita L, Altucci L, Mita DG, Marino M.** Laccase treatment impairs bisphenol A-induced cancer cell proliferation affecting estrogen receptor alpha-dependent rapid signals. *IUBMB Life* 60: 843–852, 2008.
12. **Brigelius-Flohé R, Kelly FJ, Salonen JT, Neuzil J, Zingg JM, Azzi A.** The European perspective on vitamin E: current knowledge and future research. *Am J Clin Nutr* 76: 703–716, 2002.
13. **Combs GF.** *The Vitamins: Fundamental Aspects in Nutrition and Health*. San Diego, CA: Academic, 1992.
14. **Conte C, Floridi A, Aisa C, Piroddi M, Floridi A, Galli F.** Gamma-tocotrienol metabolism and antiproliferative effect in prostate cancer cells. *Ann NY Acad Sci* 1031: 391–394, 2004.
15. **Eitsuka T, Nakagawa K, Miyazawa T.** Down-regulation of telomerase activity in DLD-1 human colorectal adenocarcinoma cells by tocotrienol. *Biochem Biophys Res Commun* 348: 170–175, 2006.
16. **Evans RM, Bishop KS.** On the existence of a hitherto unrecognized dietary factor essential for reproduction. *Science* 56: 650–651, 1922.
17. **Galluzzo P, Caiazza F, Moreno S, Marino M.** Role of ERbeta palmitoylation in the inhibition of human colon cancer cell proliferation. *Endocr Relat Cancer* 14: 153–167, 2007.
18. **Gaub MP, Bellard M, Scheuer I, Chambon P, Sassone-Corsi P.** Activation of the ovalbumin gene by the estrogen receptor involves the fos-jun complex. *Cell* 63: 1267–1276, 1990.
19. **Gooch JL, Yee D.** Strain-specific differences in formation of apoptotic DNA ladders in MCF-7 breast cancer cells. *Cancer Lett* 144: 31–37, 1999.
20. **Guthrie N, Gapor A, Chambers AF, Carroll KK.** Inhibition of proliferation of estrogen receptor-negative MDA-MB-435 and -positive MCF-7 human breast cancer cells by palm oil tocotrienols and tamoxifen, alone and in combination. *J Nutr* 127: 544S–548S, 1997.
21. **Hall JM, Couse JF, Korach KS.** The multifaceted mechanisms of estradiol and estrogen receptor signaling. *J Biol Chem* 276: 36869–36872, 2001.
22. **Hall JM, McDonnell DP.** Coregulators in nuclear estrogen receptor action: from concept to therapeutic targeting. *Mol Interv* 5: 343–357, 2005.
23. **He L, Mo H, Hadisusilo S, Qureshi AA, Elson CE.** Isoprenoids suppress the growth of murine B16 melanomas in vitro and in vivo. *J Nutr* 127: 668–674, 1997.
24. **Hoque A, Albanes D, Lippman SM, Spitz MR, Taylor PR, Klein EA, Thompson IM, Goodman P, Stanford JL, Crowley JJ, Coltman CA, Santella RM.** Molecular epidemiologic studies within the Selenium and Vitamin E Cancer Prevention Trial (SELECT). *Cancer Causes Control* 12: 627–633, 2001.
25. **Jensen EV, Cheng G, Palmieri C, Saji S, Makela S, Van Noorden S, Wahlstrom T, Warner M, Coombes RC, Gustafsson JA.** Estrogen receptors and proliferation markers in primary and recurrent breast cancer. *Proc Natl Acad Sci USA* 98: 15197–15202, 2001.
26. **Kimbro KS, Duschene K, Willard M, Moore JA, Freeman S.** A novel gene STYK1/NOK is upregulated in estrogen receptor-alpha negative estrogen receptor-beta positive breast cancer cells following estrogen treatment. *Mol Biol Rep* 35: 23–27, 2008.
27. **Kline K, Yu W, Sanders BG.** Vitamin E: mechanisms of action as tumor cell growth inhibitors. *J Nutr* 131: 161S–163S, 2001.
28. **Kuiper GG, Carlsson B, Grandien K, Enmark E, Haggblad J, Nilsson S, Gustafsson JA.** Comparison of the ligand binding specificity and transcript tissue distribution of estrogen receptors alpha and beta. *Endocrinology* 138: 863–870, 1997.
29. **Liu T, Bauskin AR, Zaunders J, Brown DA, Pankhurst S, Russell PJ, Breit SN.** Macrophage inhibitory cytokine 1 reduces cell adhesion and induces apoptosis in prostate cancer cells. *Cancer Res* 63: 5034–5040, 2003.
30. **Liu Z, Yu X, Shaikh ZA.** Rapid activation of ERK1/2 and AKT in human breast cancer cells by cadmium. *Toxicol Appl Pharmacol* 228: 286–294, 2008.
31. **Mak P, Leung YK, Tang WY, Harwood C, Ho SM.** Apigenin suppresses cancer cell growth through ERbeta. *Neoplasia* 8: 896–904, 2006.
32. **Malayer JR, Cheng J, Woods VM.** Estrogen responses in bovine fetal uterine cells involve pathways directed by both estrogen response element and activator protein-1. *Biol Reprod* 60: 1204–1210, 1999.
33. **Manas ES, Xu ZB, Unwalla RJ, Somers WS.** Understanding the selectivity of genistein for human estrogen receptor-beta using X-ray crystallography and computational methods. *Structure* 12: 2197–2207, 2004.
34. **Matthews J, Gustafsson JA.** Estrogen signaling: a subtle balance between ER alpha and ER beta. *Mol Interv* 3: 281–292, 2003.
35. **McIntyre BS, Briski KP, Gapor A, Sylvester PW.** Antiproliferative and apoptotic effects of tocopherols and tocotrienols on preneoplastic and neoplastic mouse mammary epithelial cells. *Proc Soc Exp Biol Med* 224: 292–301, 2000.
36. **Mustad VA, Smith CA, Ruey PP, Edens NK, DeMichele SJ.** Supplementation with 3 compositionally different tocotrienol supplements does not improve cardiovascular disease risk factors in men and women with hypercholesterolemia. *Am J Clin Nutr* 76: 1237–1243, 2002.
37. **Nesaretnam K, Ambra R, Selvaduray KR, Radhakrishnan A, Canali R, Virgili F.** Tocotrienol-rich fraction from palm oil and gene expression in human breast cancer cells. *Ann NY Acad Sci* 1031: 143–157, 2004.
38. **Nesaretnam K, Ambra R, Selvaduray KR, Radhakrishnan A, Reimann K, Razak G, Virgili F.** Tocotrienol-rich fraction from palm oil affects gene expression in tumors resulting from MCF-7 cell inoculation in athymic mice. *Lipids* 39: 459–467, 2004.

39. **Nesaretnam K, Dorasamy S, Darbre PD.** Tocotrienols inhibit growth of ZR-75-1 breast cancer cells. *Int J Food Sci Nutr* 51 *Suppl*: S95–S103, 2000.
40. **Nesaretnam K, Guthrie N, Chambers AF, Carroll KK.** Effect of tocotrienols on the growth of a human breast cancer cell line in culture. *Lipids* 30: 1139–1143, 1995.
41. **Nesaretnam K, Stephen R, Dils R, Darbre P.** Tocotrienols inhibit the growth of human breast cancer cells irrespective of estrogen receptor status. *Lipids* 33: 461–469, 1998.
42. **Ngah WZ, Jarien Z, San MM, Marzuki A, Top GM, Shamaan NA, Kadir KA.** Effect of tocotrienols on hepatocarcinogenesis induced by 2-acetylaminofluorene in rats. *Am J Clin Nutr* 53: 1076S–1081S, 1991.
43. **Ni J, Chen M, Zhang Y, Li R, Huang J, Yeh S.** Vitamin E succinate inhibits human prostate cancer cell growth via modulating cell cycle regulatory machinery. *Biochem Biophys Res Commun* 300: 357–363, 2003.
44. **Nilsson S, Makela S, Treuter E, Tujague M, Thomsen J, Andersson G, Enmark E, Pettersson K, Warner M, Gustafsson JA.** Mechanisms of estrogen action. *Physiol Rev* 81: 1535–1565, 2001.
45. **Norman BH, Dodge JA, Richardson TI, Borromeo PS, Lugar CW, Jones SA, Chen K, Wang Y, Durst GL, Barr RJ, Montrose-Rafizadeh C, Osborne HE, Amos RM, Guo S, Boodhoo A, Krishnan V.** Benzopyrans are selective estrogen receptor beta agonists with novel activity in models of benign prostatic hyperplasia. *J Med Chem* 49: 6155–6157, 2006.
46. **O'Lone R, Frith MC, Karlsson EK, Hansen U.** Genomic targets of nuclear estrogen receptors. *Mol Endocrinol* 18: 1859–1875, 2004.
47. **Ogawa S, Inoue S, Watanabe T, Orimo A, Hosoi T, Ouchi Y, Muramatsu M.** Molecular cloning and characterization of human estrogen receptor betacx: a potential inhibitor ofestrogen action in human. *Nucleic Acids Res* 26: 3505–3512, 1998.
48. **Ong AS.** Natural sources of tocotrienols. In: *Vitamin E in Health and Disease*, edited by Packer L and Fuchs J. New York: Marcel Dekker, 1992, p. 3–8.
49. **Packer L.** Protective role of vitamin E in biological systems. *Am J Clin Nutr* 53: 1050S–1055S, 1991.
50. **Paech K, Webb P, Kuiper GG, Nilsson S, Gustafsson J, Kushner PJ, Scanlan TS.** Differential ligand activation of estrogen receptors ERalpha and ERbeta at AP1 sites. *Science* 277: 1508–1510, 1997.
51. **Pearce ST, Jordan VC.** The biological role of estrogen receptors alpha and beta in cancer. *Crit Rev Oncol Hematol* 50: 3–22, 2004.
52. **Pfaffl MW.** A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* 29: e45, 2001.
53. **Rasool AH, Yuen KH, Yusoff K, Wong AR, Rahman AR.** Dose dependent elevation of plasma tocotrienol levels and its effect on arterial compliance, plasma total antioxidant status, and lipid profile in healthy humans supplemented with tocotrienol rich vitamin E. *J Nutr Sci Vitaminol (Tokyo)* 52: 473–478, 2006.
54. **Ricciarelli R, Zingg JM, Azzi A.** Vitamin E: protective role of a Janus molecule. *FASEB J* 15: 2314–2325, 2001.
55. **Sakai M, Okabe M, Tachibana H, Yamada K.** Apoptosis induction by gamma-tocotrienol in human hepatoma Hep3B cells. *J Nutr Biochem* 17: 672–676, 2006.
56. **Schaffer S, Muller WE, Eckert GP.** Tocotrienols: constitutional effects in aging and disease. *J Nutr* 135: 151–154, 2005.
57. **Schwenke DC.** Does lack of tocopherols and tocotrienols put women at increased risk of breast cancer? *J Nutr Biochem* 13: 2–20, 2002.
58. **Sen CK, Khanna S, Roy S.** Tocotrienols: Vitamin E beyond tocopherols. *Life Sci* 78: 2088–2098, 2006.
59. **Serbinova E, Kagan V, Han D, Packer L.** Free radical recycling and intramembrane mobility in the antioxidant properties of alpha-tocopherol and alpha-tocotrienol. *Free Radic Biol Med* 10: 263–275, 1991.
60. **Shah S, Gapor A, Sylvester PW.** Role of caspase-8 activation in mediating vitamin E-induced apoptosis in murine mammary cancer cells. *Nutr Cancer* 45: 236–246, 2003.
61. **Shah SJ, Sylvester PW.** Gamma-tocotrienol inhibits neoplastic mammary epithelial cell proliferation by decreasing Akt and nuclear factor kappaB activity. *Exp Biol Med (Maywood)* 230: 235–241, 2005.
62. **Sontag TJ, Parker RS.** Influence of major structural features of tocopherols and tocotrienols on their omega-oxidation by tocopherol-omega-hydroxylase. *J Lipid Res* 48: 1090–1098, 2007.
63. **Srivastava JK, Gupta S.** Tocotrienol-rich fraction of palm oil induces cell cycle arrest and apoptosis selectively in human prostate cancer cells. *Biochem Biophys Res Commun* 346: 447–453, 2006.
64. **Sun W, Wang Q, Chen B, Liu J, Liu H, Xu W.** Gamma-tocotrienol-induced apoptosis in human gastric cancer SGC-7901 cells is associated with a suppression in mitogen-activated protein kinase signalling. *Br J Nutr* 99: 1247–1254, 2008.
65. **Sundram K, Gapor A.** Vitamin E from palm oil: its extraction and nutritional properties. *Lipid Technol*: 137–141, 1992.
66. **Suzuki YJ, Tsuchiya M, Wassall SR, Choo YM, Govil G, Kagan VE, Packer L.** Structural and dynamic membrane properties of alpha-tocopherol and alpha-tocotrienol: implication to the molecular mechanism of their antioxidant potency. *Biochemistry* 32: 10692–10699, 1993.
67. **Sylvester PW, Nachnani A, Shah S, Briski KP.** Role of GTP-binding proteins in reversing the antiproliferative effects of tocotrienols in preneoplastic mammary epithelial cells. *Asia Pac J Clin Nutr* 11, *Suppl* 7: S452–S459, 2002.
68. **Sylvester PW, Shah S.** Intracellular mechanisms mediating tocotrienol-induced apoptosis in neoplastic mammary epithelial cells. *Asia Pac J Clin Nutr* 14: 366–373, 2005.
69. **Sylvester PW, Shah SJ.** Mechanisms mediating the antiproliferative and apoptotic effects of vitamin E in mammary cancer cells. *Front Biosci* 10: 699–709, 2005.
70. **Vladusic EA, Hornby AE, Guerra-Vladusic FK, Lakins J, Lupu R.** Expression and regulation of estrogen receptor beta in human breast tumors and cell lines. *Oncol Rep* 7: 157–167, 2000.
71. **Wada S, Satomi Y, Murakoshi M, Noguchi N, Yoshikawa T, Nishino H.** Tumor suppressive effects of tocotrienol in vivo and in vitro. *Cancer Lett* 229: 181–191, 2005.
72. **Wang S, Wang G, Barton BE, Murphy TF, Huang HF.** Beneficial effects of vitamin E in sperm functions in the rat after spinal cord injury. *J Androl* 28: 334–341, 2007.
73. **Wang Y, Chirgadze NY, Briggs SL, Khan S, Jensen EV, Burris TP.** A second binding site for hydroxytamoxifen within the coactivator-binding groove of estrogen receptor beta. *Proc Natl Acad Sci USA* 103: 9908–9911, 2006.
74. **Weiser H, Vecchi M, Schlachter M.** Stereoisomers of alpha-tocopheryl acetate. IV. USP units and alpha-tocopherol equivalents of all-rac-, 2-ambo- and RRR-alpha-tocopherol evaluated by simultaneous determination of resorption-gestation, myopathy and liver storage capacity in rats. *Int J Vitam Nutr Res* 56: 45–56, 1986.
75. **Witte D, Chirala M, Younes A, Li Y, Younes M.** Estrogen receptor beta is expressed in human colorectal adenocarcinoma. *Hum Pathol* 32: 940–944, 2001.
76. **Yap SP, Yuen KH, Lim AB.** Influence of route of administration on the absorption and disposition of alpha-, gamma- and delta-tocotrienols in rats. *J Pharm Pharmacol* 55: 53–58, 2003.
77. **Yu W, Simmons-Menchaca M, Gapor A, Sanders BG, Kline K.** Induction of apoptosis in human breast cancer cells by tocopherols and tocotrienols. *Nutr Cancer* 33: 26–32, 1999.
78. **Zhang Y, Ni J, Messing EM, Chang E, Yang CR, Yeh S.** Vitamin E succinate inhibits the function of androgen receptor and the expression of prostate-specific antigen in prostate cancer cells. *Proc Natl Acad Sci USA* 99: 7408–7413, 2002.

Paper IV: Tocotrienol activity in MCF7 breast cancer cells: involvement of Erβ signal transduction

Mol. Nutr. Food Res. 2010, *54*, 669–678 DOI 10.1002/mnfr.200900383

RESEARCH ARTICLE

Tocotrienols activity in MCF-7 breast cancer cells: Involvement of ERβ signal transduction

Raffaella Comitato[1]*, *Guido Leoni*[1]*, *Raffaella Canali*[1], *Roberto Ambra*[1], *Kalanithi Nesaretnam*[2] *and Fabio Virgili*[1]

[1] National Research Institute for Food and Nutrition, Rome, Italy
[2] Malaysian Palm Oil Board, Persiaran Institusi, Bandar Baru Bangi, Selangor, Malaysia

Received: August 7, 2009
Revised: February 12, 2010
Accepted: February 15, 2010

The term Vitamin E is utilized to describe eight molecules, subdivided into two groups, tocopherols and tocotrienols (TTs). It has been shown that specific TTs affect the growth of several lines of tumour cells, and that this activity is not shared by tocopherols. In agreement with these observations, a TTs-rich fraction from palm oil (PTRF) was reported to inhibit proliferation and induce apoptosis in several cancer cells. However, the molecular mechanism involved in TTs activity is still unclear. We have recently proposed that TTs pro-apoptotic activity involves estrogen receptor beta (ERβ) signalling. In this study, we report that, in MCF-7 breast cancer cell, expressing both ERα and ERβ, PTRF treatment increases ERβ nuclear translocation, as demonstrated by immunofluorescence experiments and significantly inhibits ERα expression (−458.91-fold of change) and complete disappearing of the protein from the nucleus. Moreover, PTRF treatment induces ER-dependent genes expression (macrophage inhibitory cytokine-1, early growth response-1 and Cathepsin D) which is inhibited by the ER inhibitor, ICI 182.780, and induces DNA fragmentation. Finally, cDNA-array experiments suggest that the activation of specific pathways in cells treated with γ-TT with respect to α-TT. Our data suggest a novel potential molecular mechanism for TTs activity.

Keywords:
Apoptosis / Estrogen nuclear receptor / Frizzled 1 / Macrophage inhibitory cytokine-1 / Tocopherols

1 Introduction

The term "Vitamin E" usually refers to a family composed of α-, β-, γ-, and δ-tocopherols and corresponding four tocotrienols (TTs). The biological role of vitamin E has been initially referred to its ability to affect the resorption-gestation performance in rodents but, so far, the real biological functions in humans have not been fully understood, and have been generally attributed to a non-specific antioxidant activity [1]. More recently, non-antioxidant activities have been reported for both tocopherols (TOCOs) and TTs with a special concern to their specific ability to affect gene expression and cell response [2].

Recent studies have suggested that TTs have specific functions and activities, distinct from those identified for TOCOs. TTs have been reported to suppress the enzymatic activity responsible for cholesterol synthesis in the liver [3], and to lower both total cholesterol and the LDL fraction. α-TT has been reported to be specifically able to prevent neuro-degenerative processes by regulating specific mediators of cell death [4]. Moreover, it has been demonstrated that oral supplementation of TTs protects against stroke and

Correspondence: Dr. Raffaella Comitato, National Research Institute for Food and Nutrition (INRAN), via Ardeatina 546, 00178 Rome, Italy
E-mail: comitato@inran.it
Fax: +390651494550

Abbreviations: Bmp-4, bone morphogenetic protein 4; **EGR-1,** early growth response-1; **ER,** estrogen receptor; **ERE,** estrogen-responsive element; **MCM,** minichromosome maintenance; **MIC-1,** macrophage inhibitory cytokine-1; **PTRF,** Tocotrienol-rich-fraction from palm oil extract; **SERCA,** sarcoplasmic/endoplasmic reticulum calcium ATPase; **SFRP,** secreted fizzled-related protein; **TOCO,** tocopherol; **TT,** tocotrienol

*These authors have contributed equally to this work.

this phenomenon was linked to their ability to reach brain tissues [5]. Previous reports by others and by our laboratory indicated that TTs can suppress cell growth and induce apoptosis in several cell lines of murine and human cancer [6, 7].

Although the molecular mechanisms underlying these potentially beneficial effects of TTs are still poorly known, some studies performed in our laboratories based on cDNA-array methods, revealed that a TT-rich fraction from palm oil (PTRF) induces a significant reduction of cell proliferation both *in vitro* in cultured breast cancer cells [8, 9] and *in vivo* in tumours induced in athymic mice by the inoculation of human breast cancer cells [10]. This approach led us to identify a set of genes transcriptionally modulated by PTRF and involved in cell cycle control. On the basis of *in silico* and *in vitro* binding experiments coupled with cell culture studies, we have more recently suggested that the effects of specific TT (γ- and δ-TT forms) on gene expression is at least in part mediated by the ability to bind to estrogen receptor beta (ERβ) in cultured MDA-MB-231 cells [7], a cell line expressing ERβ but not ERα. Such interaction results in the nuclear translocation of ERβ and the activation of specific genes containing an estrogen-responsive element (ERE) in their promoter. Finally, in agreement with other authors [11], we showed a pro-apoptotic activity of TTs in the same cell line by observing caspase-3 activation and genomic DNA fragmentation [7].

A number of cell-based and animal models have conclusively demonstrated that an intricate interplay among ERα and ERβ activities exist. ERα and ERβ are products of different genes and exhibit tissue- and cell-type specific expression and have specific roles in estrogen-dependent action *in vivo* [12, 13]. Several tissues express both receptors and when co-expressed, usually ERβ exhibits an inhibitory action on ERα-mediated gene expression and in many instances opposes the actions of ERα. The molecular mechanisms regulating the relative expression of both ERs and their direct or indirect interactions to determine each other's function have been proposed to play an important role in different pathological processes, but are still largely unknown [14].

The aim of this study was to investigate the effect of the PTRF, in a cell line expressing both ER isoforms, the MCF-7 cell line, where ERs mediated cell signaling is the result of a fine tuning and balance between ERα and ERβ activity. Moreover, cDNA-array experimental approach, allowed the identification of novel candidate pathways possibly involved in the modulation of cellular signaling by specific TT forms.

2 Materials and methods

2.1 Chemicals

PTRF was obtained from Sime Darby Plantation (Malaysia) and purified as described previously [15]. The final purity of Vitamin E in PTRF was 95–99% and typically contained 32% α-TOCO, 25% α-TT (α-T3), 29% γ-TT (γ-T3) and 14% δ-TT (δ-T3).

Purified TTs were provided by Dr. Hiroyuki Yoshimura of the Eisai Food and Chemical (Tokyo, Japan). Purity was 99% near for all TTs. Pure α-TOCO was purchased by Sigma-Aldrich (St. Louis, MO, USA). Stock solutions of PTRF and TTs were stored at −20°C in aliquots and diluted to the desired concentration in DMSO.

The non-specific ER antagonist ICI 182.780 was purchased from Tocris (Ballwin, MO, USA).

2.2 Cells lines and treatments

MCF-7 human breast cancer cells were obtained from the American Tissue Culture Collection (Manassas, VA, USA). Cells were grown in RPMI 1640 medium (Sigma-Aldrich) supplemented with 10% foetal bovine serum (Sigma-Aldrich), Pen/Strep (Invitrogen Life Science, CA, USA), 2 mM glutamine (Sigma-Aldrich) and 10% non-essential aminoacid (MEM, Sigma-Aldrich).

Before any experimental session, cells were synchronized in G_1/G_0 by starvation in serum-free medium for 3 days. Once synchronized, 5.0×10^5 cells were seeded onto multi-well plates in phenol red-free RPMI 1640 and, where appropriate, incubated with ICI 182.780 (10^{-5} M in ethanol) for 30 min.

PTRF or α-TOCO was added to the culture medium for 24 and 48 h or only 24 h, depending on the type of experiment.

Final PTRF concentration in culture media was set at 8 μg/mL, to standardize this study according with our previous reports [8]. α-TOCO concentration was 2.56 μg/mL. Moreover, concentrations of purified TTs are 2 μg/mL (β-TT) and 2.32 μg/mL (γ-TT). Concentrations reported for TTs and α-TOCO are related to the percentage present in the PTRF mixture. Concentrations of TTs used in the experiments were in the micromolar order, which is a concentration that can be achieved in the human serum, as described previously [16]. Control cells were treated with the same volumes of DMSO and/or ethanol vehicle alone.

2.3 Nuclear localization of ERβ and ERα

About 1×10^5 cells were treated with PTRF or α-TOCO, in the Lab-Tek Chamber Slides™ system (Nalge Nunc International, Rochester, NY, USA), for 24 h according to the previous article [7]. Cells were incubated with 20 μL of the diluted (1:20) ERβ or (1:50) ERα primary antibody (Santa Cruz Biotechnology, Santa Cruz, CA, USA), and stained with 100 μL DAPI to counterstain the nucleus. A Zeiss Axioskop II (Carl Zeiss AG, Oberkochen, Germany) microscope with appropriate filters was used. Images were collected and processed using the SPOT software.

2.4 RNA isolation and real-time PCR measurements

Total RNA was extracted from cells using TRI Reagent™ (Sigma-Aldrich), according to the manufacturer's instructions with some minor modifications [7]. Primers (Table 1) corresponding to selected genes were designed with Primer Express 2.0 (Applied Biosystems, Foster City, CA, USA). Real-time PCR were performed using the SuperScript™ Platinum® SYBR® Green One-Step kit (Invitrogen) according to the previous article [7]. The C_t values for each target and reference genes were obtained and their difference was calculated (ΔC_t). For normalization purpose, an identical set of reaction was prepared using primer specific for β-actin. Quantitative differences in the cDNA target among samples were determined using the mathematical model of Pfaffl [17] as described in the previous article [7].

2.5 Protein extraction and Western blot

Cells were lysed in RIPA buffer and samples were submitted to electrophoresis as described by Comitato *et al.* [7] and then incubated overnight at 4°C with a 1:500 dilution of rabbit early growth response-1 antibody (EGR-1; Santa Cruz Biotechnology), 1:500 Cathepsin D rabbit antibody (Santa Cruz Biotechnology), 1:250 Macrophage Inhibitory Cytokine-1 goat antibody (MIC-1; Novus Biologicals, Littleton, CO, USA), 1:500 bone morphogenetic protein 4 (Bmp-4; Santa Cruz Biotechnology), 1:400 ERα (Santa Cruz Biotechnology), 1:1000 ERβ (Santa Cruz Biotechnology) and 1:1000 α-tubulin mouse antibody (MP Biomedicals, Irvine, CA, USA). After washing with TPBS, membranes were incubated for 1 h at RT with 1:2000 goat anti-mouse or goat anti-rabbit or 1:5000 donkey anti-goat peroxidase-conjugated secondary antibodies (Santa Cruz Biotechnology). Specific spots were detected by chemiluminescence reagents ECL Plus (Amersham Pharmacia Biotech, Piscataway, NJ) and visualized by autoradiography by high-performance chemiluminescence film (Amersham Biosciences, Buckinghamshire, UK).

2.6 DNA laddering

DNA fragmentation was assessed according to Gooch and Yee [18] in cells after 24 h treatment with PTRF and α-TOCO. Following the isolation, DNA was electrophoresed in 1.5% agarose gels containing ethidium bromide and visualized by UVIpro Bronze acquisition system (UVITEC, Cambridge, UK).

2.7 RNA extraction and cDNA-array experiment

After 24 h of TTs treatment, total RNA was extracted from cells using RNeasy mini kit (Qiagen, Hilden, Germany), according to the manufacturer's instructions. The isolated RNA samples were sent to ServiceXS BV (Leiden, The Netherlands) where they were processed according to Affymetrix protocols.

Briefly, RNA concentration was determined by absorbance at 260 nm, and quality and integrity was verified using the Agilent 2100 Bioanalyzer (Agilent Technologies). Next, 2 μg of high quality total RNA was used with the Affymetrix Eukaryotic One-Cycle Target Labeling and Control reagents to generate biotin-labelled anti-sense cRNA. The quality of the cRNA was checked using the Agilent 2100 bio-analyzer.

The labelled cRNA was hybridized to the NuGO Affymetrix Human Genechip NuGO_Hs1a520180 (custom designed by the European Nutrigenomics Organization NuGO, consisting of 23 941 probesets including 71 control probesets, for details see http://blog.bigcat.unimaas.nl/~martijn/NuGO/).

Cell intensity File (*.cel) for each GeneChip processed are generated using Command Console Software. Three biological replicates were generated for each experimental condition. Microarrays statistical analysis was performed using oneChannelGUI R package [19], using a custom CDF file for NuGO_Hs1a520180 (based on Entrez Gene, version 10.0.0; available *via* http://nugo-r.bioinformatics.nl/NuGOR.html).

Raw signal intensity were normalized using Express method with robust multi array as background correction and keeping the remaining values to their default values.

Differentially expressed genes analysis was performed using "compute linear model fit" function in oneChannelGUI and then computing the "contrasts" with the analogous function of oneChannelGUI inherited from affylmGUI [20]. The differentially expressed genes are chosen setting as threshold a delta fold change at least 1 and with 0.05 *p*-value after Benjamini–Hochberg correction.

This approach provided expression values for 2770 (γ-TT *versus* Control) and 1168 (α-TT *versus* Control) genes differentially expressed, respectively.

Table 1. List of genes considered, GenBank identification code and sequence of primers utilized for real-time PCR

Gene	GenBank	Forward	Reverse
MIC-1	NM_004864	5′TGGTGCTCATTCAAAAGACCG3′	5′GTGGAAGGACCAGGACTGCTC3′
EGR-1	NM_001964	5′CTCCACAGGGCTTTCGGAC3′	5′GAGAGGGAGGACTTGGCTCTG3′
Cathepsin D	NM_001909	5′CTGTGAGGCCATTGTGGACAC3′	5′CAGCTTGTAGCCTTTGCCTCC3′
Bmp-4	NM_001202	5′GCCGTCATTCCGGACTACAT3′	5′GGGCGCTCAGGATACTCAAG3′
β-Actin	NM_001614	5′AGAAGGATTCCTATGTGGGGG3′	5′CATGTCGTCCCAGTTGGTGAC3′

2.8 Network analysis

A global network for all the genes differentially expressed in our set of comparison was built using MIMI plug-in [21] of Cytoscape [22], selecting as query parameter: input gene and nearest neighbourhood. This resulted in a network of around 6000 genes on which we have mapped microarray expression data for retrieve only the subset of genes differentially expressed in one comparison as described previously.

Over-representation of KEGG pathway was computed using DAVID [23] web server (http://david.abcc.ncifcrf.gov/home.jsp). The selected pathways are chosen according to a *p*-value below 0.05.

2.9 Statistics and data presentation

All data are presented as the means ± SE of at least three independent experiments. Statistical analysis was performed with "R software" from "R Foundation for Statistical Computing" (Vienna, Austria). Real-time data were analyzed by one-way ANOVA with repeated measures followed by the Bartlett and Fligner–Killeen test for homogeneity of variance. Dunnett *post hoc* test was used to evaluate difference among multiple conditions. About $p \leq 0.05$ were considered to be statistically significant.

Figures and tables present one out of at least three independent experiments providing similar results or the mean (± SE) of at least three experiments, respectively.

3 Results and discussion

TTs are attracting increasing scientific interest as candidate for specific biological actions, well beyond their antioxidant activity, affecting cellular function and survival.

PTRF is a standardized mixture obtained from palm oil. It contains about 70% w/w of TTs together with α-TOCO and carotenoids and it has been often utilized as a tool to study the biological effects of TTs. We initially considered the effect of PTRF in inhibiting breast cancer cell growth [24, 25]. In these early reports, the observation that PTRF pro-apoptotic activity in breast cancer cells was not associated with a specific ERs expression profile, led to the conclusion that PTRF activity was independent of estrogen-related signaling. However, we have recently reported a direct interaction of PTRF and purified TTs with ERβ, by combining results obtained with software-based docking simulations, *in vitro* estrogen (E2) displacement assays and immunocytochemistry and by evaluating ERE-dependent gene expression in the MDA-MB-231 cell line, expressing only Erβ [7].

As mentioned in Section 1, an intricate interplay between ERα and ERβ activities has been demonstrated in a number of experimental models. However, the final effect of the ligand–receptor interaction in the presence of the co-expression of both ER isoforms is not obvious in different circumstances, and warrants further investigation. ERα and ERβ belong to the nuclear receptors superfamily and more specifically to the family of steroid receptors acting as ligand-regulated transcription factors [26]. There are several complex relationships between the two receptors. On a simplification attempt, such complex relationship has been considered as a "*yin/yang*" like interaction. In fact, ERα and ERβ have different biological functions, as indicated by their different expression patterns [14]. In particular, as mentioned above, ERβ appears to act as a dominant regulator of estrogen signalling, and when co-expressed with ERα, it causes a concentration-dependent reduction in ERα-mediated transcriptional activation [12, 27]. However, it has been argued that the ERβ-dependent antagonism of ERα-mediated responses does not represent a general mechanism in ER signalling, as it could be restricted to a limited number of genes [28]. Therefore, on the basis of these observations, it is possible to speculate that TTs binding to ERβ, modulate a specific cell response that also results in the down-regulation of ERβ expression inducing a shift of ERα/ERβ balance in favour of ERβ, which can be responsible for the reduction of MCF-7 tumourigenicity [13]. Consistent with this view, E2 has been reported to increase cell proliferation and to cause tumour formation in MCF-7 cells expressing higher levels of ERα than ERβ [14].

A number of evidences suggest that there are several distinct pathways by which estrogens, *via* ERs, can regulate a wide spectrum of biological processes [29]. In the classical model of ER action, ligand-activated ERs specifically bind DNA at EREs through their DNA-binding domains and bring co-regulators to the transcription start site. In order to identify which pathways were activated, we tested the expression of a selected small set of ERE-containing genes (namely, MIC-1, EGR-1 and Cathepsin D) in MCF-7 cells, at baseline and after PTRF treatment, in the presence of ICI 182.780, a specific ERs inhibitor. This set of genes was selected from a database previously identified using a gene array approach [10]; moreover, they were also considered and analyzed in our previous article on MDA-MB-231 cells [7].

Therefore, the aim of this study was to investigate the molecular pathways activated by treatment with either PTFR or specific TTs, leading to apoptosis in breast tumour cells expressing both ER isoforms. To this purpose, we have utilized MCF-7 human breast cancer cells, expressing ERα and ERβ at the level of both RNA transcript and protein [30].

Our first aim was to confirm our previous observations of an agonist interaction of TTs with ERβ. In agreement with the previous reports [31], in baseline conditions, the localization of ERβ in MCF-7 cells is predominantly cytoplasmatic (Fig. 1). After 24 h of treatment with PTRF, an evident nuclear staining is visible, indicating a significant translocation of the receptor inside the nucleus. On the other hand, at the same time point, ERα which is normally located in the

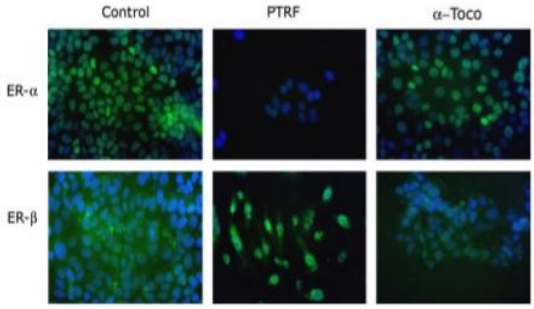


Figure 1. Cells were treated for 24 h with PTRF (8 μg/mL) or α-TOCO (2.56 μg/mL). PTRF treatment was associated with ERβ strong nuclear staining and total disappearance of nuclear ERα. On the other hand, no difference was observed in cells treated with α-TOCO in respect to control. One out of the three independent experiments yielding similar results is shown.

nucleus [32], is no more visible. Moreover, in agreement with our previous results obtained on MDA-MB-231 cells [7], α-TOCO treatment of MCF-7 is not associated with any significant changes of ERα and ERβ localization, with respect to control cells. These results indicate that PTRF, but not α-TOCO, induces the nuclear transfer of ERβ also in the presence of ERα, suggesting a specific role of ERβ as mediator of PTFR effects. Moreover, in cells treated with γ-TT and δ-TT, but not α-TT, we have observed the same trend of localization associated to PTRF of ERα and ERβ (data not shown).

In MCF-7 cells, PTRF treatment induces a significant increase of the expression of MIC-1 and EGR-1 mRNAs, after 24 and 48 h of treatments (Figs. 2A and B, respectively). Unexpectedly, Cathepsin D mRNA is down-regulated after 24 h and slightly up-regulated after 48 h of PTRF treatment. Pre-treatment with ICI 182.780 partially prevents the modulation of all genes. As expected, the expression of an ERE-lacking gene, Bmp-4, utilized as a "negative control" to exclude the possibility of a non-specific activation of gene expression by PTRF, remains unchanged in all the experimental conditions.

Protein levels are consistent with the amount of mRNA measured: MIC-1 and EGR-1 levels were significantly increased by PTRF treatment (Fig. 3) while remained stable at the baseline level in the presence of ICI 182.780 pre-treatment (Fig. 3). Surprisingly, and in disagreement with mRNA levels, Cathepsin D protein expression increased. According to the unchanged mRNA expression, the protein levels of the ERE-lacking gene Bmp-4 were not affected by treatments. ICI 182.780 treatment alone was not associated to any effects.

Our data indicate that the effects of PTRF on gene expression modulation in MCF-7 cells are due, at least in part, to its interaction with ERβ transcriptional pathway. ICI 182.780 is a well-known specific estrogen antagonist blocking the majority of pathways mediated by ERα and ERβ [33]. The inhibition of the effects of PTRF in the presence of this inhibitor, as demonstrated by Real-time PCR and Western blot analyses, clearly indicates that ERβ specifically mediates PTRF effects.

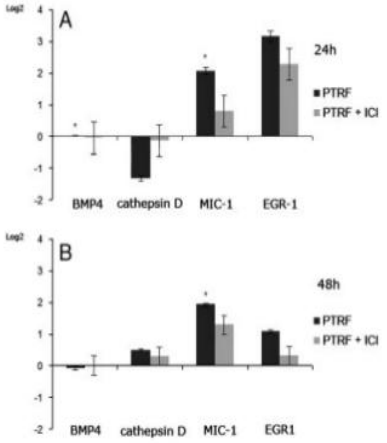


Figure 2. Expression of Bmp-4, EGR-1, MIC-1 and Cathepsin D genes in MCF-7 cells treated for 24 and 48 h with 8 μg/mL PTRF. Where appropriate, cells were pre-treated with ICI 182.780 for 30 min. Gene expression was analyzed by real-time quantitative PCR and results were log transformed (logarithm 2) in order to obtain data symmetrically distributed. Statistical significance was calculated by Dunnett *post hoc* test ($p = * \leq 0.05$).

Moreover, the treatment with PTRF is associated with alterations of the morphology of MCF-7 cells. We observed a loosening of cellular spindle shaped morphology with cells becoming smaller and rounded, detaching and flattening. Trypan blue staining indicates that the majorities of floating cells are not necrotic and still alive (data not shown), suggesting that the cells might have started apoptosis. In order to determine whether PTRF may induce apoptosis, we extracted DNA from adherent and non-adherent cells and then subjected it to agarose gel electrophoresis. Figure 4 shows that DNA fragmentation occurs in MCF-7 cells treated with PTRF. This effect is in agreement with previously

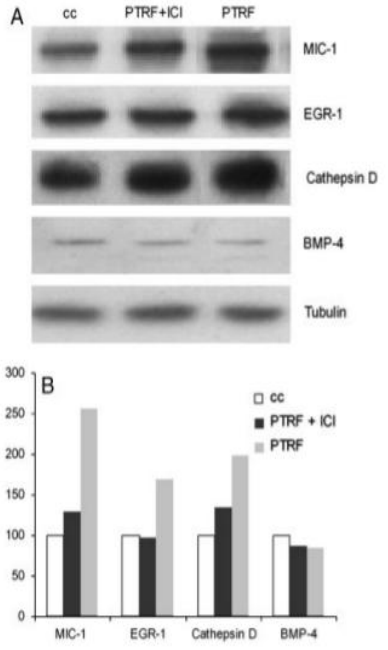


Figure 3. Total proteins from MCF-7 cells treated for 48 h with 8 μg/mL PTRF. Where appropriate, cells were pre-treated with ICI 182.780 for 30 min.

published articles by other authors that showed evident pro-apoptotic [34–36] and inhibitory effects on cell growth [24, 37] either by purified TTs or by PTRF.

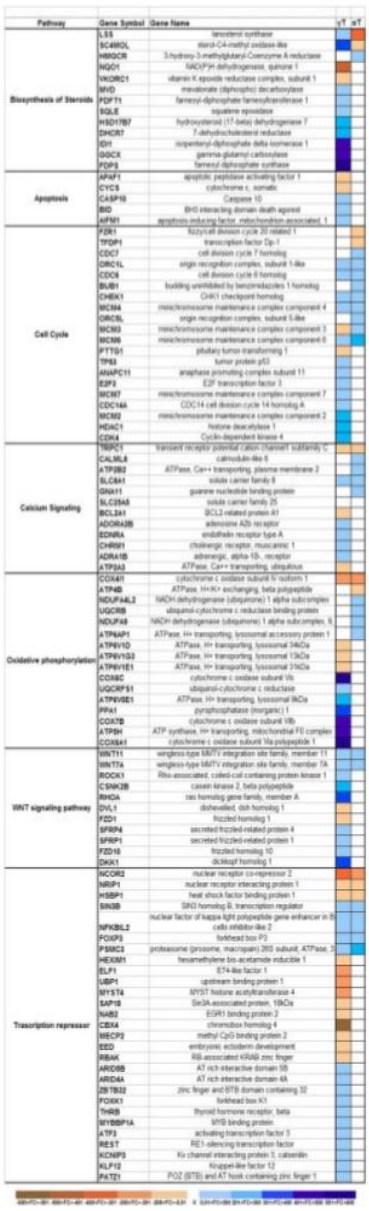
In our previous study on MDA-MB-231 breast cancer cells, we have showed that the activity of PTRF mainly relied on γ- and δ-TT, but not on α-TT and α-TOCO [7]. We therefore hypothesized that γ- and δ-TT could be the real effectors of PTRF activity also in MCF-7. In order to corroborate this hypothesis, cDNA-array-based experiments were performed, utilizing RNA extracted from MCF-7 cells treated with purified TTs. This approach confirmed the Real-time PCR experiments and, in addition, showed a strong down-regulation of ERα (ESR1) mRNA expression, in agreement with immunocytochemistry assay. Moreover, this approach allowed the identification of specific cellular signalling networks associated with TTs effects. Among others, the network analysis performed comparing γ-TT and α-TT treatments *versus* control samples revealed the activation of specific molecular pathway, in cells treated with γ-TT but not with α-TT (Fig. 5). In particular, we observed a significant up-regulation of mRNAs encoding for enzymes involved in the steroid biosynthesis pathway, and the modulation of several

Figure 4. DNA laddering in MCF-7 cells treated for 24 h with PTRF (8 μg/mL) or α-TOCO (2.56 μg/mL) and then DNA was extracted and electrophoresed in 1.5% agarose gels containing ethidium bromide and visualized by UVIpro Bronze acquisition system (UVITEC). The typical apoptotic laddering was observed only in PTRF-treated cells.

repressors of transcriptional factors: namely the up-regulation of forkhead box K1, forkhead box P3, retinoblastoma-binding protein 1, ARID domain-containing protein 5B, and NF-kappa-B inhibitor-like protein 2 and the down-regulation of E74-like factor 1, hexamethylene bis-acetamide-inducible protein 1, methyl CpG-binding protein 2, heat shock factor binding protein 1 and nuclear receptor co-repressor 2. The modulation of the expression of these genes has been associated to the development of mammary carcinoma [38, 39], suggesting that they play a role in the adaptability of cancer cells to pro-apoptotic stimuli.

Using a mini plug-in of Cytoscape software, we built up the network of the specific genes regulated by PTFR. The network analysis revealed that some of the repressors of the transcription factors, NRIP1, THR, hexamethylene bis-acetamide-inducible protein 1 and methyl CpG-binding protein 2, directly modulate ERα mRNA expression [39–42], suggesting a role for these molecules in the down-regulation of ERα observed in MCF-7 treated with γ-TT.

cDNA arrays also indicate that Wnt genes are affected by TT treatment. Wnts regulate a variety of cellular activities, including proliferation, migration, polarity and gene expression [43] and they have been proposed to be implicated in tumour formation in different organs [44]. Interestingly, we observed a strong decrease of frizzled-1 mRNA, one of the main seven trans-membrane receptors of Wnt molecules [45]. The expression of frizzled-1 receptors has been reported to be usually up-regulated in breast cancer [46]. Moreover, we found an up-regulation of genes encoding for secreted frizzled-related protein-1 and -4 (SFRP-1



and SFRP-4). SFRPs represent a family of extracellular factors that antagonizes Wnt activies [47]. An up-regulation of SRFP1 and SFRP4 has been associated to apoptosis [48, 49] whereas their down-regulation has been observed in breast cancer [49]. It has been recently reported that the activation of ER-independent Wnt/β-catenin signalling by estrogen in uterine epithelial cells during the early growth phase, significantly contributes to the ER-dependent late growth response [50]. Moreover, Yokota *et al.* identified SFRP1 as an estrogen-inducible gene in stromal cells [51].

At single gene level, we observed a strong down-regulation of MCM3 gene, one of the members of the hexameric minichromosome maintenance (MCM) complex. An incomplete functionality of the hexameric MCM complex has been reported to prevent the unwinding of DNA double helix during the S-phase of mitosis [52]. The protein encoded by this gene is involved in the initiation of eukaryotic genome replication and is usually over-expressed in several types of tumour [53]. Moreover, we detected a modulation of cell cycle pathway following γ-TT treatment. In particular, we observed a down-regulation of pituitary tumour-transforming gene together with an up-regulation of tumour protein p53 gene. Similarly to MCM3, pituitary tumour-transforming gene is usually over-expressed in breast cancer cells [54] and encodes for human securin, a protein that interacts with p53 blocking its interaction with DNA and thus inhibiting its ability to induce cell death [55].

Moreover, focusing to the Ca^{2+}-related signalling pathways, we observed a strong down-regulation of the expression of BCL2A1, a negative modulator of mitochondrial apoptosis, and of ATP2A3 gene, which encodes for a sarcoplasmic/endoplasmic reticulum calcium ATPase (SERCA)-3 responsible for resorption of cytosolic calcium inside the endoplasmatic reticulum [56]. It is well known that high levels of cytosolic Ca^{2+} act as a mitochondrial pro-apoptotic signal. In fact, in normal conditions the equilibrium between cytosolic and endoplasmatic reticulum calcium levels is maintained by SERCA pumps combined with proteins Bax/Bcl2 balance modulating the release of cytocrome c from mitochondria. Accordingly, low expression of SERCA and Bcl2 result in high level of cytosolic Ca^{2+} [57].

◀ **Figure 5.** Over-represented KEGG pathways and GO biological processes in γ- and α-TTs *versus* control cDNA arrays experiments. From all the expressed genes, according to array analysis, were chosen only genes with a delta FC value (expr(treatment)-expr(Control)) at least 1 and a *p*-value adjusted for Benjamini–Hochberg correction of 0.05. The entire set of differentially expressed genes in treatments *versus* control was used as input for DAVID web server. Only the biological processes with a *p*-value below 0.05 were considered as over-represented. We also illustrated delta FC values for treatments *versus* controls (see text for more details).

Finally, as demonstrated by immunocytochemistry experiments, γ-TT but not α-TT, activates apoptosis in MCF-7. cDNA-array data confirm that α-TT does not activate any pro-apoptotic gene. On the other hand, γ-TT down-regulates cytocrome *c* gene expression and activates the apoptosis inducing factor–mitochondrion associated. Moreover, γ-TT modulates different genes involved in mitochondrial phosphorylation, including different subunits of the cythocrome *c* oxidase. In particular, the subunit IV (COX4I1) is down-regulated by 276 folds in comparison to control. The decrease of the subunit IV has been demonstrated to disassembly the enzyme complex and to compromise membrane potential, leading to a decrease of ATP levels and to a sensitization of the cells to apoptosis [58]. Therefore, on the basis of these observations, we can hypothesize that the molecular mechanism underlying γ-TT induced apoptosis in MCF-7 cells is, at least in part, mitochondria-driven.

In general, we remark that the biological processes specifically modulated by γ-TT, but not α-TT, share many similarities with gene expression profiling of estrogen-regulated genes in MCF-7 breast cancer cells after treatment with anti-estrogens drugs [59].

On the basis of these results and previously published data [7], we can hypothesize that γ-TT induction of specific metabolic pathways and of apoptosis act through an ERβ-associated signal transduction pathway. However, the involvement of different transduction pathways in TTs effects cannot be excluded. Further studies are needed to fully understand the molecular mechanisms underlying TT effects.

4 Concluding remarks

These data confirm our previous original observation indicating an estrogenic activity of specific TTs forms and corroborate the notion that TTs biological activity is specific, distinct from TOCOs and not "restricted" to their antioxidant capacity [60, 61]. Moreover, the evidence that TT do not share a significant number of biological activities with the related TOCOs, also suggests that the possibility to pull them apart from the "heterogeneous" family of Vitamin E, and to propose new roles and functions for this class of molecules in human health and disease.

Moreover our observations may have a pharmacological concern. In the last years, the estrogenic therapy has been mainly based on the targeting of ERα and ERβ; and in particular, a number of studies addressed a protective role of ERβ against breast cancer development. Therefore, the detection of molecules acting as specific target for ERα and ERβ would open up to original "nutritional management" opportunities and possibly to novel therapeutic approaches.

All these observation strongly suggest that the molecules of nutritional interest eliciting ERβ activity can act as possible tumour suppressors. Our observations could open new avenues for a specific role of TTs in regulating gene expression and in modulating the growth of breast cancer cell and of other tumour in part or totally dependent on estrogen signals.

The authors thank Dr. Hiroyuki Yoshimura of the Eisai Food and Chemical Co. for the generous gifts of purified TTs. Moreover, we wish to thank Professor Maria Marino for helpful discussions and comments on the manuscript. This work was supported by the Malaysian Palm Oil Board, and the Ministry of Education, University, by the Research of Italy (FISR "Safe-eat") and Ministry of Agricultural, Nutritional Policies and Forestry (MiPAAF-NUME).

The authors have declared no conflict of interest.

5 References

[1] Traber, M. G., Atkinson, J., Vitamin E, antioxidant and nothing more. *Free Radic. Biol. Med.* 2007, *43*, 4–15.

[2] Azzi, A., Gysin, R., Kempna, P., Munteanu, A. *et al.*, Vitamin E mediates cell signaling and regulation of gene expression. *Ann. NY Acad. Sci.* 2004, *1031*, 86–95.

[3] Pearce, B. C., Parker, R. A., Deason, M. E., Qureshi, A. A. *et al.*, Hypocholesterolemic activity of synthetic and natural tocotrienols. *J. Med. Chem.* 1992, *35*, 3595–3606.

[4] Khanna, S., Roy, S., Parinandi, N. L., Maurer, M. *et al.*, Characterization of the potent neuroprotective properties of the natural vitamin E alpha-tocotrienol. *J. Neurochem.* 2006, *98*, 1474–1486.

[5] Khanna, S., Patel, V., Rink, C., Roy, S. *et al.*, Delivery of orally supplemented alpha-tocotrienol to vital organs of rats and tocopherol-transport protein deficient mice. *Free Radic. Biol. Med.* 2005, *39*, 1310–1319.

[6] Kline, K., Yu, W., Sanders, B. G., Vitamin E and breast cancer. *J. Nutr.* 2004, *134*, 3458S–3462S.

[7] Comitato, R., Nesaretnam, K., Leoni, G., Ambra, R. *et al.*, A novel mechanism of natural vitamin E tocotrienol activity: involvement of ERbeta signal transduction. *Am. J. Physiol. Endocrinol. Metab.* 2009, *297*, E427–E437.

[8] Nesaretnam, K., Ambra, R., Selvaduray, K. R., Radhakrishnan, A. *et al.*, Tocotrienol-rich fraction from palm oil and gene expression in human breast cancer cells. *Ann. NY Acad. Sci.* 2004, *1031*, 143–157.

[9] Nesaretnam, K., Dorasamy, S., Darbre, P. D., Tocotrienols inhibit growth of ZR-75-1 breast cancer cells. *Int. J. Food Sci. Nutr.* 2000, *51*, S95–S103.

[10] Nesaretnam, K., Ambra, R., Selvaduray, K. R., Radhakrishnan, A. *et al.*, Tocotrienol-rich fraction from palm oil affects gene expression in tumors resulting from MCF-7 cell inoculation in athymic mice. *Lipids* 2004, *39*, 459–467.

[11] Xu, W. L., Liu, J. R., Liu, H. K., Qi, G. Y. *et al.*, Inhibition of proliferation and induction of apoptosis by gamma-tocotrienol in human colon carcinoma HT-29 cells. *Nutrition* 2009, *25*, 555–566.

[12] Liu, M. M., Albanese, C., Anderson, C. M., Hilty, K. *et al.*, Opposing action of estrogen receptors alpha and beta on cyclin D1 gene expression. *J. Biol. Chem.* 2002, *277*, 24353–24360.

[13] Chang, E. C., Frasor, J., Komm, B., Katzenellenbogen, B. S., Impact of estrogen receptor beta on gene networks regulated by estrogen receptor alpha in breast cancer cells. *Endocrinology* 2006, *147*, 4831–4842.

[14] Matthews, J., Gustafsson, J. A., Estrogen signaling: a subtle balance between ER alpha and ER beta. *Mol. Interv.* 2003, *3*, 281–292.

[15] Sundram, K., Gapor, A., Vitamin E from palm oil: its extraction and nutritional properties. *Lipid Technol.* 1992, *a*, 137–141.

[16] Rasool, A. H., Yuen, K. H., Yusoff, K., Wong, A. R. *et al.*, Dose dependent elevation of plasma tocotrienol levels and its effect on arterial compliance, plasma total antioxidant status, and lipid profile in healthy humans supplemented with tocotrienol rich vitamin E. *J. Nutr. Sci. Vitaminol. (Tokyo)* 2006, *52*, 473–478.

[17] Pfaffl, M. W., A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* 2001, *29*, e45.

[18] Gooch, J. L., Yee, D., Strain-specific differences in formation of apoptotic DNA ladders in MCF-7 breast cancer cells. *Cancer Lett.* 1999, *144*, 31–37.

[19] Sanges, R., Cordero, F., Calogero, R. A., oneChannelGUI: a graphical interface to Bioconductor tools, designed for life scientists who are not familiar with R language. *Bioinformatics* 2007, *23*, 3406–3408.

[20] Wettenhall, J. M., Simpson, K. M., Satterley, K., Smyth, G. K., affylmGUI: a graphical user interface for linear modeling of single channel microarray data. *Bioinformatics* 2006, *22*, 897–899.

[21] Gao, J., Ade, A. S., Tarcea, V. G., Weymouth, T. E. *et al.*, Integrating and annotating the interactome using the MiMI plugin for cytoscape. *Bioinformatics* 2009, *25*, 137–138.

[22] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S. *et al.*, Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003, *13*, 2498–2504.

[23] Dennis, G., Jr., Sherman, B. T., Hosack, D. A., Yang, J. *et al.*, DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.* 2003, *4*, P3.

[24] Nesaretnam, K., Stephen, R., Dils, R., Darbre, P., Tocotrienols inhibit the growth of human breast cancer cells irrespective of estrogen receptor status. *Lipids* 1998, *33*, 461–469.

[25] Guthrie, N., Gapor, A., Chambers, A. F., Carroll, K. K., Inhibition of proliferation of estrogen receptor-negative MDA-MB-435 and -positive MCF-7 human breast cancer cells by palm oil tocotrienols and tamoxifen, alone and in combination. *J. Nutr.* 1997, *127*, 544S–548S.

[26] Beato, M., Gene regulation by steroid hormones. *Cell* 1989, *56*, 335–344.

[27] Pettersson, K., Delaunay, F., Gustafsson, J. A., Estrogen receptor beta acts as a dominant regulator of estrogen signaling. *Oncogene* 2000, *19*, 4970–4978.

[28] Rissman, E. F., Roles of oestrogen receptors alpha and beta in behavioural neuroendocrinology: beyond Yin/Yang. *J. Neuroendocrinol.* 2008, *20*, 873–879.

[29] Hall, J. M., Couse, J. F., Korach, K. S., The multifaceted mechanisms of estradiol and estrogen receptor signaling. *J. Biol. Chem.* 2001, *276*, 36869–36872.

[30] Vladusic, E. A., Hornby, A. E., Guerra-Vladusic, F. K., Lakins, J. *et al.*, Expression and regulation of estrogen receptor beta in human breast tumors and cell lines. *Oncol. Rep.* 2000, *7*, 157–167.

[31] Witte, D., Chirala, M., Younes, A., Li, Y. *et al.*, Estrogen receptor beta is expressed in human colorectal adenocarcinoma. *Hum. Pathol.* 2001, *32*, 940–944.

[32] Catalano, S., Marsico, S., Giordano, C., Mauro, L. *et al.*, Leptin enhances, *via* AP-1, expression of aromatase in the MCF-7 cell line. *J. Biol. Chem.* 2003, *278*, 28668–28676.

[33] Malayer, J. R., Cheng, J., Woods, V. M., Estrogen responses in bovine fetal uterine cells involve pathways directed by both estrogen response element and activator protein-1. *Biol. Reprod.* 1999, *60*, 1204–1210.

[34] Ahn, K. S., Sethi, G., Krishnan, K., Aggarwal, B. B., Gamma-tocotrienol inhibits nuclear factor-kappaB signaling pathway through inhibition of receptor-interacting protein and TAK1 leading to suppression of antiapoptotic gene products and potentiation of apoptosis. *J. Biol. Chem.* 2007, *282*, 809–820.

[35] Sakai, M., Okabe, M., Tachibana, H., Yamada, K., Apoptosis induction by gamma-tocotrienol in human hepatoma Hep3B cells. *J. Nutr. Biochem.* 2006, *17*, 672–676.

[36] Shah, S. J., Sylvester, P. W., Gamma-tocotrienol inhibits neoplastic mammary epithelial cell proliferation by decreasing Akt and nuclear factor kappaB activity. *Exp. Biol. Med. (Maywood)* 2005, *230*, 235–241.

[37] Nesaretnam, K., Guthrie, N., Chambers, A. F., Carroll, K. K., Effect of tocotrienols on the growth of a human breast cancer cell line in culture. *Lipids* 1995, *30*, 1139–1143.

[38] Pilarsky, C., Wenzig, M., Specht, T., Saeger, H. D. *et al.*, Identification and validation of commonly overexpressed genes in solid tumors by comparison of microarray data. *Neoplasia* 2004, *6*, 744–750.

[39] Wittmann, B. M., Fujinaga, K., Deng, H., Ogba, N. *et al.*, The breast cell growth inhibitor, estrogen down regulated gene 1, modulates a novel functional interaction between estrogen receptor alpha and transcriptional elongation factor cyclin T1. *Oncogene* 2005, *24*, 5576–5588.

[40] Graupner, G., Zhang, X. K., Tzukerman, M., Wills, K. *et al.*, Thyroid hormone receptors repress estrogen receptor activation of a TRE. *Mol. Endocrinol.* 1991, *5*, 365–372.

[41] Sharma, D., Blum, J., Yang, X., Beaulieu, N. *et al.*, Release of methyl CpG binding proteins and histone deacetylase 1 from the estrogen receptor alpha (ER) promoter upon reactivation in ER-negative human breast cancer cells. *Mol. Endocrinol.* 2005, *19*, 1740–1751.

[42] Cavailles, V., Dauvois, S., L'Horset, F., Lopez, G. *et al.*, Nuclear factor RIP140 modulates transcriptional activation by the estrogen receptor. *EMBO J.* 1995, *14*, 3741–3751.

[43] Moon, R. T., Shah, K., Developmental biology: signalling polarity. *Nature* 2002, *417*, 239–240.

[44] Brennan, K. R., Brown, A. M., Wnt proteins in mammary development and cancer. *J. Mammary Gland Biol. Neoplasia* 2004, *9*, 119–131.

[45] He, X., Semenov, M., Tamai, K., Zeng, X., LDL receptor-related proteins 5 and 6 in Wnt/beta-catenin signaling: arrows point the way. *Development* 2004, *131*, 1663–1677.

[46] Milovanovic, T., Planutis, K., Nguyen, A., Marsh, J. L. *et al.*, Expression of Wnt genes and frizzled 1 and 2 receptors in normal breast epithelium and infiltrating breast carcinoma. *Int. J. Oncol.* 2004, *25*, 1337–1342.

[47] Glinka, A., Wu, W., Delius, H., Monaghan, A. P. *et al.*, Dickkopf-1 is a member of a new family of secreted proteins and functions in head induction. *Nature* 1998, *391*, 357–362.

[48] Melkonyan, H. S., Chang, W. C., Shapiro, J. P., Mahadevappa, M. *et al.*, SARPs: a family of secreted apoptosis-related proteins. *Proc. Natl. Acad. Sci. USA* 1997, *94*, 13636–13641.

[49] Zhou, Z., Wang, J., Han, X., Zhou, J. *et al.*, Up-regulation of human secreted frizzled homolog in apoptosis and its down-regulation in breast tumors. *Int. J. Cancer* 1998, *78*, 95–99.

[50] Hou, X., Tan, Y., Li, M., Dey, S. K. *et al.*, Canonical Wnt signaling is critical to estrogen-mediated uterine growth. *Mol. Endocrinol.* 2004, *18*, 3035–3049.

[51] Yokota, T., Oritani, K., Garrett, K. P., Kouro, T. *et al.*, Soluble frizzled-related protein 1 is estrogen inducible in bone marrow stromal cells and suppresses the earliest events in lymphopoiesis. *J. Immunol.* 2008, *181*, 6061–6072.

[52] Forsburg, S. L., Eukaryotic MCM proteins: beyond replication initiation. *Microbiol. Mol. Biol. Rev.* 2004, *68*, 109–131.

[53] Ha, S. A., Shin, S. M., Namkoong, H., Lee, H. *et al.*, Cancer-associated expression of minichromosome maintenance 3 gene in several human cancers and its involvement in tumorigenesis. *Clin. Cancer Res.* 2004, *10*, 8386–8395.

[54] Solbach, C., Roller, M., Fellbaum, C., Nicoletti, M. *et al.*, PTTG mRNA expression in primary breast cancer: a prognostic marker for lymph node invasion and tumor recurrence. *Breast* 2004, *13*, 80–81.

[55] Bernal, J. A., Luna, R., Espina, A., Lazaro, I. *et al.*, Human securin interacts with p53 and modulates p53-mediated transcriptional activity and apoptosis. *Nat. Genet.* 2002, *32*, 306–311.

[56] Hovnanian, A., SERCA pumps and human diseases. *Subcell. Biochem.* 2007, *45*, 337–363.

[57] Demaurex, N., Distelhorst, C., Cell biology. Apoptosis – the calcium connection. *Science* 2003, *300*, 65–67.

[58] Li, Y., Park, J. S., Deng, J. H., Bai, Y., Cytochrome c oxidase subunit IV is essential for assembly and respiratory function of the enzyme complex. *J. Bioenerg. Biomembr.* 2006, *38*, 283–291.

[59] Musgrove, E. A., Sergio, C. M., Loi, S., Inman, C. K. *et al.*, Identification of functional networks of estrogen- and c-Myc-responsive genes and their relationship to response to tamoxifen therapy in breast cancer. *PLoS One* 2008, *3*, e2987.

[60] Virgili, F., Marino, M., Regulation of cellular signals from nutritional molecules: a specific role for phytochemicals, beyond antioxidant activity. *Free Radic. Biol. Med.* 2008, *45*, 1205–1216.

[61] Zingg, J. M., Vitamin E: an overview of major research directions. *Mol. Aspects Med.* 2007, *28*, 400–422.

5 References

1 Force T. et al; **cardio toxicity of kinase inhibitors: the prediction and translation of preclinical models to clinical outcomes**. Nat Rev Drug Discov. 2011 Feb;10(2):111-26.

2 Lomenick B et al; **Identification of direct protein targets of small molecules.** ACS Chem Biol. 2011 Jan 21;6(1):34-46. Epub 2010 Nov 30

3 Carlson SM, White FM; **Using small molecules and chemical genetics to interrogate signaling networks.**ACS Chem Biol. 2011 Jan 21;6(1):75-85. Epub 2010 Nov 29.

4 Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B; **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** Nat Methods 2008, 5(7):621-628.

5 Sultan M et al.;**A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.**Science 2008, 321(5891):956-960

6 Wang ET et al.:**Alternative isoform regulation in human tissue transcriptomes.**Nature 2008, 456(7221):470-476

7 Unwin RD et al.;**Quantitative proteomics reveals posttranslational control as a regulatory factor in primary hematopoietic stem cells.** Blood 2006,107:4687-4694.

8 Tian Q et al.;**Integrated genomic and proteomic analyses of gene expression in Mammalian cells.**Mol Cell Proteomics 2004, 3:960-969.

9 Chen G et al-; **Discordantprotein and mRNA expression in lung adenocarcinomas.**Mol Cell Proteomics 2002, 1:304-313.

10 Likic VA et al.; **Systems biology: the next frontier for bioinformatics.**Adv Bioinformatics 2010:268925. Epub 2011 Feb 9

11 LI Q. et al.; **A network-based multi-target computational estimation scheme for anticoagulant activities of compounds.**PLoS One. 2011 Mar 22;6(3):e14774.

12 Monji H et al**.;Interaction site prediction by structural similarity to neighboring clusters in protein-protein interaction networks**.BMC Bioinformatics. 2011 Feb 15;12 Suppl 1:S39**.**

13 Minai R et al.;**Method for comparing the structures of protein ligand-binding sites and application for predicting protein-drug interactions**.Proteins. 2008 Jul;72(1):367-81**.**

14 DeKeyser JG;**Selective phthalate activation of naturally occurring human constitutive androstane receptor splice variants and the pregnane x receptor.**Toxicol Sci. 2011 Apr;120(2):381-91. Epub 2011 Jan 12.

15 Jänicke RU; **The do's and don'ts of p53 isoforms**. Biol Chem. 2009 Oct;390(10):951-63**.**

16 Gingeras T.;**Origin of phenotypes: Genes and transcripts**Genome Res. 2007. 17: 682-690

17 Burge, C.; Karlin, S.; **Prediction of complete gene structures in human genomic DNA**. Journal of Molecular Biology 1997 268 (1): 78–94

18 Salzberg SL, Delcher AL, Kasif S, White O.;**Microbial gene identification using interpolated Markov models**Nucleic Acids Res. 1998 Jan 15;26(2):544-8.

19 van Baren MJ, Koebbe BC, Brent MR**Using N-SCAN or TWINSCAN to predict gene structures in genomic DNA sequences**. Curr Protoc Bioinformatics. 2007 Dec;Chapter 4:Unit 4.8

20 YosephBarash, John A. Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J. Blencowe, and Brendan J. Frey.; **Deciphering the Splicing Code.** Nature, 465:7294, May 6, 2010

21 Martin DI .;**Transcriptional enhancers--on/off gene regulation as an adaptation to silencing in higher eukaryotic nuclei**. Trends Genet. 2001 Aug;17(8):444-8

22 Pozzoli U., Sironi M.;**Silencers regulate both constitutive and alternative splicing events in mammals**. Cell Mol Life Sci. 2005 Jul;62(14):1579-604.

23 Kuhn EJ, Geyer PK**.; Genomic insulators: connecting properties to mechanism**. Curr Opin Cell Biol. 2003 Jun;15(3):259-65.

24 Geserick C, Meyer HA, Haendler B.;**The role of DNA response elements as allosteric modulators of steroid receptor function.**Mol Cell Endocrinol. 2005 May 31;236(1-2):1-7.

25 Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach; **Global quantification of mammalian gene expression control.** M Nature. 2011 May 19;473(7347):337-42-

26 Ross J**; mRNA stability in mammalian cells**. Microbiol Rev. 1995 Sep;59(3):423-50.
27 't Hoen PA, Hirsch M, de Meijer EJ, de Menezes RX, van Ommen GJ, den Dunnen; **mRNA degradation controls differentiation state-dependent differences in transcript and splice variant abundance.**JTNucleic Acids Res. 2011 Jan;39(2):556-66. Epub 2010 Sep 17..
28 Neu-Yilik G, Amthor B, Gehring NH, Bahri S, Paidassi H, Hentze MW, Kulozik AE;**Mechanism of escape from nonsense-mediated mRNA decay of human beta-globin transcripts with nonsense mutations in the first exon**RNA. 2011 May;17(5):843-54.
29 de Lima Morais DA, Harrison PM.; **Large-scale evidence for conservation of NMD candidature across mammals** PLoS One. 2010 Jul 21;5(7):e11695
30 Tang F, Lao K, Surani MA. **Development and applications of single-cell transcriptome analysis** Nat Methods. 2011 Apr;8(4 Suppl):S6-11
31 Shannon P, Markiel A, Ozier O, et al.**Cytoscape: a software environment for integrated models of biomolecular interaction networks**. Genome Res. 13 (11): 2498–504.
32 Stein A., Ceol A., Aloy P.; **3did: identification and classification of domain-based interactions of known three-dimensional structure** Nucleic Acids Res. 2011, 39, D718-723
33 de Vries SJ, van Dijk M, Bonvin AM.**The HADDOCK web server for data-drivenbiomolecular docking** Nat Protoc. 2010;5(5):883-97. Epub 2010 Apr 15
34 Cosconati S, Forli S, Perryman AL, Harris R, Goodsell DS, Olson AJ **Virtual Screening with AutoDock: Theory and Practice**. Expert Opin Drug Discov. 2010 Jun 1;5(6):597-607
35 Hegyi H, Kalmar L, Horvath T, Tompa P;**Verification of alternative splicing variants based on domain integrity, truncation length and intrinsic protein disorder**. Nucleic Acids Res. 2011 Mar;39(4):1208-19. Epub 2010 Oct 23
36 R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunesekaran, G. Ceric, K. Forslund, L. Holm, E.L. Sonnhammer, S.R. Eddy, A. Bateman;**The Pfam protein families database:** Nucleic Acids Research (2010) Database Issue 38:D211-222
37 Sigrist CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N.; **PROSITE, a protein domain database for functional characterization and annotation.**Nucleic Acids Res. 38(Database issue)161-6 (2010
38 Puntervoll P, Linding R, Gemünd C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, Ferrè F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Küster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ**; ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins**. Nucleic Acids Res. 2003 Jul 1;31(13):3625-3
39 Attwood TK**.; The PRINTS database: a resource for identification of protein families.**BriefBioinform. 2002 Sep;3(3):252-63.
40 Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995**). SCOP: a structural classification of proteins database for the investigation of sequences and structures**. J. Mol. Biol. 247, 536-540
41 Catherine Bru, Emmanuel Courcelle, SébastienCarrère, YoannBeausse, Sandrine Dalmar, and Daniel Kahn The **ProDom database of protein domain families: more emphasis on 3D**. Nucleic Acids Res. (2005) 33: D212-D215
42 R.D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, P. Gunesekaran, G. Ceric, K. Forslund, L. Holm, E.L. Sonnhammer, S.R. Eddy, A. Bateman;**The Pfam protein families database**; Nucleic Acids Research (2010) Database Issue 38:D211-222
43 Tress ML et al.;**The implications of alternative splicing in the ENCODE protein complement**. Proc NatlAcad Ski U S A. 2007 Mar 27;104(13):5495-500. Epub 2007 Mar 19
44 Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R.**GENCODE: producing a reference annotation for ENCODE**. Genome Biol. 2006;7Suppl 1:S4.1-9. Epub 2006 Aug 7.
45 Birzele F, Csaba G, Zimmer R.;**Alternative splicing and protein structure evolution.**Nucleic Acids Res.2008;**36**:550–558
46 Melamud E, Moult J. **Structural implication of splicing stochastics**. Nucleic Acids Res. 2009;**37**:4862–4872
47 Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying**

mammalian transcriptomes by RNA-Seq. Nat Methods 2008, 5(7):621-628.
48 Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D et al**: A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome**. Science 2008, 321(5891):956-960.
49 Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. Nature 2008, 456(7221):470-476
50 Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ**: Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing**. Nat Genet 2008, 40(12):1413-1415.
51 Krug K, Nahnsen S, Macek**B Mass spectrometry at the interface of proteomics and genomics**. Mol Biosyst. 2011 Feb;7(2):284-91. Epub 2010 Oct 21.
52 Tanner S, Shen Z, Ng J, Florea L, Guigo R, Briggs SP, Bafna V: **Improving gene annotation using peptide mass spectrometry**. Genome Res 2007, 17:231-239.
53 Tress ML, Bodenmiller B, Aebersold R, Valencia A**: Proteomics studies confirm the presence of alternative protein isoforms on a large scale**. Genome Biol 2008, 9:R162.
54 Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I et al**: The SWISS-PROT protein knowledgebase and its supplement TrEMBL** 2003. Nucleic Acids Res 2003, 31(1):365-370
55 Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, Gelfand MS, Sunyaev S: **Increase of functional diversity by alternative splicing**. Trends Genet 2003, 19(3):124-128.
56 Birzele F, Küffner R, Meier F, Oefinger F, Potthast C, Zimmer R. **ProSAS: a database for analyzing alternative splicing in the context of protein structures**. Nucleic Acids Res. 2008 Jan;36(Database issue): D63-8. Epub 2007 Oct 11.
57 Shionyu M, Yamaguchi A, Shinoda K, Takahashi K, Go M. **AS-ALPS: a database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse**. Nucleic Acids Res. 2009 Jan;37(Database issue):D305-9. Epub 2008 Nov 10
58 Brigelius-Flohe, B; Traber (1999).**"Vitamin E: function and metabolism**". FASEB 13: 1145–1155
59 Pearce, B. C., Parker, R. A., Deason, M. E., Qureshi, A. A. et al.,**Hypocholesterolemic activity of synthetic and natural tocotrienols**. J. Med. Chem. 1992, 35, 3595–3606.
60 Khanna, S., Roy, S., Parinandi, N. L., Maurer, M. et al.,Characterization **of the potent neuroprotective properties of the natural vitamin E**. J. Neurochem. 2006, 98, 1474–1486
61 Kline K, Yu W, and Sanders BG. **Vitamin E: mechanisms of action as tumor cell growth inhibitors**. The Journal of nutrition 131: 161S-163S, 2001.
62 McIntyre BS, Briski KP, Gapor A, and Sylvester PW. **Antiproliferative and apoptotic effects of tocopherols and tocotrienols on preneoplastic and neoplastic mouse mammary epithelial cells**. Proceedings of the Society for Experimental Biology and Medicine Society for Experimental Biology and Medicine611 292-301, 2000.
63 Sylvester PW, Wali VB, Bachawal SV, Shirode AB, Ayoub NM, Akl MR **Tocotrienol combination therapy results in synergistic anticancer response** Front Biosci. 2011 Jun 1;17:3183-95
64 Virgili F. **Tocotrienol-rich fraction from palm oil and gene expression in human breast cancer cells**. Annals of the New York Academy of Sciences 1031: 143-157,2004.
65 Nesaretnam K, Ambra R, Selvaduray KR, Radhakrishnan A, Reimann K, Razak G, and Virgili F. Tocotrienol**-rich fraction from palm oil affects gene expression in tumors resulting from MCF-7 cell inoculation in athymic mice**. Lipids 39: 459-467, 2004
66 Chang, E. C., Frasor, J., Komm, B., Katzenellenbogen, B. S.,**Impact of estrogen receptor beta on gene networks regulatedby estrogen receptor alpha in breast cancer cells**.Endocrinology 2006, 147, 4831–4842.
67 Matthews, J., Gustafsson, J. A**., Estrogen signaling: a subtle balance between ER alpha and ER beta**. Mol. Interv. 2003, 3, 281–292.
68 Rissman, E. F**., Roles of oestrogen receptors alpha and beta in behavioural neuroendocrinology**: beyond Yin/Yang. J. Neuroendocrinol. 2008
69 Demaurex, N., Distelhorst, C., **Cell biology. Apoptosis the calcium connection**. Science 2003, 300, 65–67.

70 Kisselman G, Qiu W, Romanov V, Thompson CM, Lam R, Battaile KP, Pai EF, Chirgadze NY;**X-CHIP: an integrated platform for high-throughput protein crystallization and on-the-chip X-ray diffraction data collection**. ActaCrystallogr D BiolCrystallogr. 2011 Jun;67(Pt 6):533-9. Epub 2011 May 12.

71 Ambrish Roy, AlperKucukural, Yang Zhang**. I-TASSER: a unified platform for automated protein structure and function prediction**. Nature Protocols, vol 5, 725-738 (2010)

72 Whitman S, Bang X, Shalaby R, Shtivelman **Alternatively spliced products CC3 and TC3 have opposing effects on apoptosis**. E.Mol Cell Biol. 20(2):583-93 (2010).

73 SeolD.W., Billiar T.R.**A caspase-9 variant missing the catalytic site is an endogenous inhibitor of apoptosis.**J.Biol. Chem. 274:2072-2076(1999)

74 Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, Chinappi M, Tramontano A, Bozzoni I.**A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA.** Cell. 14;147(2):358-69 (2011).

75 Rao N, Nguyen S, Ngo K, Fung-Leung WP.**A novel splice variant of interleukin-1 receptor (IL-1R)-associated kinase 1 plays a negative regulatory role in Toll/IL-1R-induced inflammatory signaling.**Mol Cell Biol. 25(15):6521-32 (2005).

76 Kishore S, Khanna A, Zhang Z, Hui J, Balwierz PJ, Stefan M, Beach C, Nicholls RD, Zavolan M, Stamm S.**The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing.**Hum Mol Genet.19(7):1153-64. (2010).