



LSHG – CT – 2003 - 503265

BioSapiens

A European Network for Integrated Genome Annotation

Network of Excellence

Life Sciences, Genomics and Biotechnology for Health

Final Activity Report

Period covered: from 1.1.2004 to 30.6.2009

Date of preparation: 14.8.2009

Start date of project: 1.1.2004

Duration: 66 months

Project coordinator name:

Janet Thornton

Project coordinator organisation name:

**European Molecular Biology
Laboratory**

BioSapiens: Final Activity Report (1.1.2004 – 30.6.2009)

A. Project execution

1.1 Summary description of project objectives	
A. Objectives	3
B. Contractors involved	3
1.2 Work performed and end results	
A. Integrating activities	4
B. Joint Research Activities	8
C. Spreading of Excellence Activities	19
D. Management Actions	23
1.3 Objectives reached	25
1.4 Methodologies and approaches	29
1.5 Relationship to state-of-the-art	30
1.6 Impact of BioSapiens on industry and research sectors	30

B. Dissemination and use

2.1 Final plans for using and disseminating knowledge	32
2.2 Intellectual property	32

Appendix 1 – Plan for using and disseminating knowledge

Section 2 – Dissemination of knowledge

Appendix 2 - List of publications

1. PROJECT EXECUTION

1.1 Summary description of project objectives

A. Objectives

The objective of the BIOSAPIENS Network of Excellence is to support a large scale, concerted effort to annotate genome data by laboratories distributed around Europe, using both informatics tools and input from experimentalists. Through integration the network aims to improve bioinformatics research in Europe, by providing a focus for annotation and by the organisation of European meetings and workshops to encourage cooperation, rather than duplication of effort. The annotations generated by the network are to be made available in the public domain and will be easily accessed on the web. Details of the project and those involved are available on the project website //http:www.biosapiens.info

B. Contractors involved

In its final year, the network consisted of the following contractors:

Prof. Janet Thornton, EMBL - European Bioinformatics Institute, UK
Prof. Dmitrij Frishman, German Nat. Centre for Environment and Health, Germany
Prof. Jacques van Helden, Université Libre de Bruxelles, Belgium
Prof. Alfonso Valencia, Centro Nacional de Investigaciones Oncológicas, Spain
- *previously Centro Nacional de Biotecnología-CSIC, Spain*
Dr Roderic Guigo, Centre for Genomic Regulation, Spain
- *previously IMM Laboratory of Computational Genomics, Spain*
Dr Tim Hubbard, Wellcome Trust Sanger Institute, UK
Prof Dr Thomas Lengauer, Max-Planck Institute für Informatik, Germany
Prof Michal Linial, The Hebrew University of Jerusalem, Israel
Prof Anna Tramontano, University of Rome "La Sapienza", Italy
Prof Gunnar von Heijne, Stockholms Universitet, Sweden
Dr Richard Mott, Wellcome Trust for Human Genetics, UK
Prof Christine Orengo and Prof David Jones, University College London, UK
Prof Gert Vriend, University of Nijmegen, The Netherlands
Dr Anne-Lise Veuthey, Swiss Institute of Bioinformatics, Switzerland
Prof Søren Brunak, Technical University of Denmark, Denmark
Prof Esko Ukkonen, University of Helsinki, Finland
Prof Stylianos Antonarakis, University of Geneva, Switzerland
Prof Laszlo Patthy, Institute of Enzymology, Hungary
Prof Dietmar Schomburg, Technical University of Braunschweig, Germany
- *previously University of Cologne, Germany*
Prof. Antoine Danchin, Institut Pasteur, France
Dr Leszek Rychlewski, BioInfoBank Institute, Poland
Dr Vincent Schachter, Genoscope-CEA, France
Prof Martin Vingron, Max-Planck Institute for Molecular Genetics, Germany
Prof Rita Casadio, University of Bologna, Italy
Dr Nikos Darzentas/ Dr Christos Ouzounis (*until end 2008*), Centre for Research and Technology Hellas, Greece (*from 1.1.2006*)

Management of the Consortium

The network is coordinated by Professor Janet Thornton (director@ebi.ac.uk) at the European Bioinformatics Institute, and managed by Dr Kerstin Nyberg (knyberg@ebi.ac.uk). The steering committee also includes Prof Anna Tramontano (Training Coordinator), Prof Alfonso Valencia (Outreach Coordinator) and Prof Soren Brunak (Thematic WP Coordinator)

1.2 Work performed and end results

This first phase of the Network (Year 1) was a proof of principle phase, during which the virtual institute was established and the infrastructure to allow distributed annotation and access to the virtual knowledge centre developed.

The second phase of the Network (Year 2) firmly established the virtual institute and the infrastructure to allow distributed annotation and access to the virtual knowledge centre. During this year, new tools were developed and a variety of Genome Annotations delivered through the BioSapiens Portal.

The third phase of the Network (Years 3 to end) continued to establish and refine the virtual institute and the infrastructure to allow distributed annotation and access to the virtual knowledge centre. During the final three and a half years many new tools were still being developed and a variety of Genome Annotations delivered through the BioSapiens Portal.

A. Integrating Activities

At the outset, the network consisted of 25 laboratories from 14 different European countries. To launch the project 37 members of the consortium, representing all contractors gathered for the first Annual General Meeting held in Rome on 26 – 27 February 2004. Individual work packages held meetings to coordinate their research activities and interactions and to discuss the work programme, including deliverables and milestones. Working together towards these goals has strengthened the community and generated a cooperative approach to genome annotation. Each work package required close contact and interaction between partners. Already in the first year, the remarkable progress was made to develop the technical integration needed to facilitate the distributed annotation effort, using the DAS technology (Distributed Annotation System). Each laboratory installed a DAS server and provided annotations as appropriate that can be viewed at any site through a DAS client. At the end of the first year, working DAS servers had been installed in 15 laboratories, each providing some test annotations. This integrated system provided the proof of principle required in the first year, that the infrastructure was appropriate to provide basic distributed annotation.

The two thematic work packages (on Down's syndrome and HCV and HIV viruses) started well. They have promoted interactions between the bioinformatics groups and the experimental biologists and allowed the experimentalists to exploit the most up-to-date annotations available.

During this first year, an extensive outreach programme included press releases, workshops to promote BioSapiens, presentations to scientists and the general public

on the importance of genome annotations and a meeting of representatives of European Master's courses in Bioinformatics. The visibility and impact of bioinformatics in Europe was improved through the many meetings and workshops BioSapiens organised and helped to support, including CAPRI: Critical Assessment of Predictions on Interactions; CASP6: Critical Assessment of Techniques for Protein Structure Prediction; Bologna Winter School in Bioinformatics; Pre-SIG meeting on 'Genome Annotation' at ISMB/ECCB2004. To foster worldwide interactions, the BioSapiens network was described at several other international meetings in Europe and US, with encouraging interactions between computational biologists and experimentalists as one of the objectives. During this period the first meeting with our scientific advisory board was held in Rome. This group comprises almost exclusively of experimentalists, and the aim of this meeting was to explain our goals and receive their advice.

The second Annual General Meeting was held in Cambridge on 11th – 12th April 2005. 37 members of the consortium, representing all contractors gathered for this meeting. Each work package had required close contact and interaction between partners during the first year, and the individual work packages took the opportunity to hold individual meetings. The progress had continued to develop the technical integration needed to facilitate the distributed annotation effort, using the DAS technology (Distributed Annotation System). At the end of the second year working DAS servers had been installed in 24 laboratories, providing 64 data sources for protein, genomic and structural annotations. The BioSapiens portal was well established and provided the central hub for the network and the central shop-window for the project's annotations, so that all the annotations from all the laboratories could be simply viewed at a single site.

The two thematic work packages (on Down's Syndrome and HCV and HIV viruses) progressed well in the second year. In accordance with the original plan these work packages had essentially been completed and a new thematic work package on annotating the ENCODE proteins agreed and to start on 1/1/2006.

The extensive outreach programme continued in the second year. Additional activities in this year were the support for the 2005 Bologna Winter School in Bioinformatics; a Pre-SIG meeting on 'Genome Annotation' at ECCB2005 in Madrid, and the organization of a bioinformatics session during the ESF bi-annual conference on "functional genomics and disease" in Oslo (September 2005), where BioSapiens also organized a workshop to discuss with a set of relevant experimental biologist their needs and views on automatic annotation. These experimentalists were very positive about BioSapiens and the DAS annotations and encouraged us to proceed with vigour.

The third Annual General Meeting was held in Barcelona on March 27 – 29, 2006. 65 members of the consortium, representing all contractors, attended the meeting. The individual work packages also held their meetings in connection with the AGM. It was agreed that in the fourth year of the network some work packages would be combined, as the need for deeper integration had become apparent. Further progress in the technical integration needed to facilitate this distributed annotation effort had been made, using the DAS technology (Distributed Annotation System). At the end of the third year, working DAS servers installed in 24 laboratories, providing 69 data sources for protein, genomic and structural annotations. The major challenge in the

third year was to improve the DAS clients, and to move towards making the information presented easier to understand and use. This involved classifying the annotations, so that they could be sensibly grouped, and adequately documented. In the longer term, this work will allow us to compare annotations of the same features derived in different laboratories, possibly generating consensus annotations and thereby enhancing the integration. The BioSapiens portal was well established and provided the central hub for the network. However additional DAS clients were being developed and the web services technology championed by the EMBRACE Network of Excellence utilised, with their collaboration.

The ENCODE thematic work package, in which many of the partners were involved, had been a major challenge of integration and delivery. The paper describing this work was accepted for publication in *Proceedings of the National Academy of Sciences*. This work certainly promoted interactions between the bioinformatics groups and will allow the experimentalists to exploit the most up-to-date annotations available. Two new thematic work packages were agreed for 2007: (1) Annotating proteins associated with cancer, and (2) Identifying responsible genes under the phenotype-associated Quantitative Trait Loci (QTLs) identified in a large mouse screen – the phenotypes included many diseases, so one goal was to identify the genes which contribute to a given disease.

The outreach programme continued during the third year, and incorporated the organisation and support of many meetings and workshops, including the 7th ‘Critical Assessment of Techniques for Protein Structure Prediction (CASP), the 2006 Bologna Winter School in Bioinformatics; a Pre-SIG meeting entitled “Genome Annotation: a BioSapiens Network of Excellence Initiative” at ECCB2006 in Eilat, and the ENCODE project.



Fig 1.2.1 WP104 meeting at Kloster Seeon

The fourth Annual General Meeting was held at Kloster Seeon, near Munich, on April 2nd to 4th, 2007. The first year of the new combined work packages had been

completed, and some success in achieving deeper integration could be noted. There had been further progress in the technical integration needed to facilitate this distributed annotation effort, and the major achievement had been to develop the sequence annotation ontology necessary to display and combine information from different laboratories. This involved classifying the annotations, so that they could be sensibly grouped, and adequately documented. This work would allow us in the final year of the network to compare annotations of the same features derived in different laboratories, allowing the generation of consensus annotations and thereby enhancing the integration. In addition the CARGO ‘widget-based’ approach had been developed, specifically for the cancer thematic project, to provide greater flexibility in presenting different types of relevant data. The Network was also utilising the web services technology championed by the EMBRACE Network of Excellence, with their collaboration.

Three thematic projects were ongoing in the fourth year. In the continuation of the ENCODE project, experimental validation of protein translation had been started using a variety of wet methods. This work was focussed on a set of alternate splice variants (ASVs) first identified in the ENCODE project, but flagged as ‘suspicious’ through the BioSapiens analysis of their protein sequences and structures. Preliminary results showed that although some of these ASVs were transcribed, they were often, but not always, present at much lower levels. The two new thematic work packages in 2007 (‘Annotating proteins associated with cancer’, and ‘Identifying responsible genes under the phenotype-associated Quantitative Trait Loci (QTLs) identified in a large mouse screen’) had made a good start, and work would continue in year 5.

The extensive outreach programme continued in the fourth year, and included workshops to promote BioSapiens (see <http://www.biosapiens.info/>). A major event was the Special Interest Group workshop on Automated Genome Annotation, which was organised prior to the ECCB/ISMB meeting in Vienna in July 2007. We continued our support for the Bologna Winter School with another “BioSapiens corner” (<http://www.biocomp.unibo.it/~school2007/>). The consortium was also involved in the production of a book, dedicated to the BioSapiens methods and developments, with Dmitrij Frishman as editor. The book was published by Springer in 2008.

The BioSapiens steering committee and consortium had played major parts in 2006 in the preparation of a proposal for the ESFRI programme. The “European Life Sciences Infrastructure for Biological Information” (ELIXIR) was successful and work started in January 2008.

The BioSapiens project was awarded a 6 month extension, so the final reporting period covers the last 18 months of the project. Over 60 members of the consortium, representing all contractors, met for the fifth and final Annual General Meeting held in Brussels, on April 3rd, 2008. The restructured work packages, defined last year, had provided a better platform for integration in this last period. There had also been further progress in the technical integration needed to facilitate this distributed annotation effort, using the DAS technology (Distributed Annotation System). Each laboratory was now providing annotations as appropriate that can be viewed at any site through various DAS clients. Following the development and publication of the sequence annotation ontology, the annotations were now presented on the portal in a

logical order, allowing comparisons between annotations, so enhancing the integration. The BioSapiens portal and the CARGO 'widget-based' approach together provide great flexibility in presenting different types of relevant data. The project was also utilizing the web services technology championed by the EMBRACE Network of Excellence, with their collaboration.

Work on the three thematic WPs continued to the end of the project. In the (WP108) ENCODE project, the construction of the Epipe pipeline to analyse the data from the complete human genome automatically had made progress. The Cancer Genome Thematic Project (WP109) has started by developing tools to handle the data and by analysing available public data to try to understand the difference between cancer associated SNPs and acceptable human mutations. In WP110 on QTL analysis, several new methods and tools had been developed to integrate genetic and expression data in mouse.

The extensive outreach programme continued in this last period, and included workshops to promote BioSapiens (see <http://www.BioSapiens.info/>). The major event was the workshop entitled "From Genome to Proteome and Biological Function" held in April 2008 at the Université Libre de Bruxelles, Brussels, which was open to all scientists and attracted well over a hundred participants. We continued our support for the Bologna Winter School with two more "BioSapiens corners" (<http://www.biocomp.unibo.it/~school2008/> and [/~school2009/](http://www.biocomp.unibo.it/~school2009/)).

The "BioSapiens book", "Modern Genome Annotation: The BioSapiens Network", edited by Dmitrij Frishman and Alfonso Valencia with chapters written by most BioSapiens partners, was published by Springer in November 2008. In addition, four review articles and a commentary, written and coordinated by the network partners were published in *Genome Biology* in early 2009.

In the last 6 months of the project, outreach efforts were focused on promoting and supporting the proposed ELIXIR infrastructure for biological information. The origins and development of this proposal, based in part on the three bioinformatics Networks of Excellence were described at the ELIXIR Stakeholders meeting in Copenhagen, May 2009. Søren Brunak's presentation, "The European Bioinformatics Area – BioSapiens, EMBRACE and ENFIN", explained how the three Networks had effectively created a new form of collaboration within the European bioinformatics community and would continue in the web services and DAS registries which were being set up.

B. Joint Research Activities

The research activities can be summarised as (i) infrastructure developments, (ii) new methods for annotation and (iii) the provision of annotations:

(i) Infrastructure Developments for Annotation and Technical Integration

During the first year, some of the infrastructure necessary for technical integration and combined visualisation of data from different laboratories was developed. The choice had been to use the sophisticated Ensembl DAS client for genome data. However a new protein DAS client had to be developed to view the protein-related data in 1D (DASTY) and to help improve a client (SPICE) developed in the Sanger

Institute for viewing these data in 3D. Prototype versions were already developed in the first year. These clients were made available in the public domain after further testing during the second period. In principle, the annotations could be viewed from any site worldwide with a DAS client, but we also designed a prototype BioSapiens DAS Portal, where all the annotations provided by the network were listed and viewed simultaneously. Additional infrastructure was also provided, including reference servers for UniProt DAS (that provide the sequences to be annotated) and links between genes and expression data, available in ArrayExpress, allowing direct access of expression profiles for a particular human gene through DAS.

During the second year of the project, a complete client-server infrastructure was developed and implemented for genome and protein annotation. Further extensions to infrastructure functionality were being continually developed.

For viewing genomic data, the Ensembl DAS client had been in widespread use for some time at the end of the second year. For viewing protein-related data in 1D, the novel DAS client DASTY¹ has been developed using Java and Macromedia Flash. For viewing protein-related data in 3D, there had been ongoing development of the SPICE² client, in collaboration with the Wellcome Trust Sanger Institute. Both the DASTY and the SPICE clients had been released to the public by the end of 2005.

In early 2006 (at the time of writing the second annual report) there were 64 data sources for protein, genomic, and structural annotations provided by the BioSapiens network partners. These annotations could be listed and viewed via the publicly-available BioSapiens DAS Portal.

The BioSapiens DAS Portal also provided access to the BioSapiens DAS Server Information Resource (BioSapiensDIR) which was developed and released as a comprehensive, centralised resource to view the registered DAS servers within the BioSapiens network. This information was continuously updated via automated health-checks, such as testing DAS requests (e.g. using the dsn, features, types and entry_points commands), and checking whether a UniProt-specific data-source supplies an MD5 checksum with returned annotation data. BioSapiensDIR is closely linked to data-source registration information stored in the Sanger public DAS registry.

During the third reporting period, the following improvements to the infrastructure for annotations were made:

Work Package	Improvements to Bioinformatics Infrastructure in 2006
2, Regulators and Promoters	<ul style="list-style-type: none"> - Tools for display and analysis of ChIP-chip data, within Ensembl - Web services for cis-regulation RSAT, including workflows
3, Expression	<ul style="list-style-type: none"> - Seamless integration of expression data with Ensembl browser - Web site for EEL Prediction
4, Variation (haplotypes and SNPs)	<ul style="list-style-type: none"> - GSCANDB developments for storing and visualising eQTLs and extension to other species

5, Protein families, orthologues	- Protocol to improve protein family annotations in DAS, by adding reliability scores or free text
6, Membrane Proteins and Ligands	- Maintenance of GPCR and KchannelDB databases for specific membrane protein families
8, Post-translation modification and localisation	- Several new Neural Network predictors
9, Sequence and Structure to Function	- Web tool to predict specificity-determining functional residues - The GOTax platform – to identify related proteins, independent of sequence similarity
10, Protein-protein complexes	- STRING extended to include annotations collected from other data resources or manually -
11, Pathways and networks	- Further development of UniPathway database of metabolic pathways -
12, Distributed Annotations for Proteins	- Test and training sets for prediction programs for 6 different sequence features - Web Service (CaPSuLo) access to predictors of subcellular location - Web-based DAS client for protein annotation (DASTY2)
13, Integration of annotations for protein structures	- XML exchange for 3D structural motifs & associated web server
21, DAS Technology	- Documentation for ProServer DAS server and client - Adapted SPICE for GENCODE - Server to link GENCODE and UniProt - Introduced DAS capabilities into JalView - Incorporated ordering of features into DASTY - Merger of DAS Registry and BioSapiensDIR

During the fourth reporting period, the following improvements to the infrastructure for annotations have been made:

Work Package	Improvements to Bioinformatics Infrastructure in 2007
101. Gene Definition/ Alternative splicing	- Development of FixPred pipeline, to improve gene identification - Wiki site for experimental pilot project to investigate translation of alternative spliced variants from GENCODE annotation
102. Gene Regulation and Expression	- Extension of Web Services and development of Workflows for regulatory sequence analysis - New interfaces for transcriptional regulator prediction - ‘Regulatory Sequence Analysis Tools’ RSAT Web Services
104 Protein Function Annotation	- Servers and databases to identify functionally important residues

	<ul style="list-style-type: none"> - Automatic clustering and annotation of protein families - Data resources for ligand binding sites - Improvements to Gene-3D website
105 Post-translation modification, membrane and localisation prediction	<ul style="list-style-type: none"> - DaRSy – Dataset Retrieval system to extract sequences of proteins with PTMs
106, Protein Complexes, networks and pathways	<ul style="list-style-type: none"> - Release 7.1 STRING database - NeAT – Network Analysis Tools website - UniPathway website and database
107 DAS Infrastructure	<ul style="list-style-type: none"> - Adding alignment DAS support - Implementing alignment reference server - Extension to provide molecular interaction data - Standardised protein feature ontology - Extension of DASTY2 DAS client & DAS registry
108 Thematic WP ENCODE	<ul style="list-style-type: none"> - Epipe – automatic pipeline to annotate splice variants - Methods to identify the principal isoform
109 Cancer	<ul style="list-style-type: none"> - Development of CARGO – Cancer and Related Genes Online - DAS server for Cancer-related genes

In the fifth and final reporting period, the following improvements to the infrastructure for annotations have been made:

Work Package	Improvements to Bioinformatics Infrastructure in 2008-9
101. Gene Definition/ Alternative splicing	Improvement of FixPred pipeline to improve gene identification Development of a pipeline to identify Selenoprotein genes
102. Gene Regulation and Expression	Server to compare expression of orthologous genes in mouse and human New interfaces for transcriptional regulator prediction Extension of Regulatory Sequence Analysis Tools (RSAT) to improve the detection of cis-acting elements in multicellular organisms
104 Protein Function Annotation	BioSerf - Protein structure prediction server Prototype server for analysing mutations LigASite – database of biologically relevant binding sites has been upgraded FireStar and FireDB have been further developed DESITE – a database of destabilising regions in proteins has been established Prototype protein-protein interface prediction server GODot for function prediction has been further developed MAISTAS – to model ASVs and assess their viability
105 Post-translation modification, membrane and localisation prediction	New servers for PTM modification & prediction of membrane topology Web service implementation of method to predict specific kinases for phosphorylation sites
106, Protein Complexes, networks and pathways	Release 8.0 STRING database <i>M. Pneumoniae</i> website DIMA (Domain Interaction Map)

107 DAS Infrastructure	Maintenance and improvements to DAS registry Further developments of SPICE – for 3D structural data DAS server for protein interaction data Improvements to DASTY2 DAS client Incorporation into IntAct as molecular viewer
108 Thematic WP ENCODE	Epipe – automatic pipeline to annotate splice variants: inclusion of new methods from BioSapiens partners
109 Thematic WP Cancer	Development of CARGO – Cancer and Related Genes Online
110 QTL Analysis	Resource for the genetic analysis of complex traits in mouse

These infrastructures covered all aspects of the network and made both the annotations and the tools easier to access. Many different changes to the DAS technology improved its usability.

(ii) New Methods for Annotation

One of the objectives of the network was to encourage the development of methods for genome and proteome annotation. In the first year of the project several such new methods were developed including *promoterwise*, to identify cis-regulatory motifs from comparative genome data; a new version of **TMHMM** to predict the topology of membrane proteins incorporating protein domain annotations; **SecretomeP** – to identify all secreted proteins and **ProP** – to identify proteins with furin or general PC cleavage sites. We further developed standards and tools to retrieve, store and display metabolic pathway data, a prototype pipeline for functional annotation from structure – **ProFunc**, and a database of predicted 3D protein models.

During 2005, various annotation methods were developed, enhanced and released. Four transmembrane helix predictors were developed and made available for general prediction use: **TMHMM2.0** which is a very fast single-sequence based method, and other methods which exploit the evolutionary information derived by multiple sequence alignment - **ENSEMBLE** (which combines two hidden Markov models and one neural network), **PRODIV_TMhmm_0.91**, and a new version of **MEMSAT**. All four methods were used to annotate the UNIPROT-human proteome. A neural-network based method (**NetAcet**) was developed for prediction of amino-terminal acetylation., and there was further development of SecretomeP to analyse the human IPI database. A new web service (**MSDmotif**) was developed for protein annotation, which facilitates defining, searching and generating statistics for small structural motifs. A relational database (**gscandb**) was developed for viewing genome-scan results and gene-annotation information by automatically linking to the genome annotation databases Ensembl, UCSC and NCBI gene. The latest release of Human And Vertebrate Analysis aNd Annotation (**HAVANA**) was used for re-annotating HSA21. ENCODE project patterns were subjected to a combination of computational predictions, manual curation and experimental verification to build a reference annotation which was used in a community experiment (**EGASP**) to evaluate computational predictions. The reannotation of metazoan Affymetrix microarrays was performed using a sequence-based procedure which exploits Biomart mappings to assign and annotate microarray design elements. Diverse annotations were combined and loaded into the ArrayExpress data warehouse (**AEDW**). The **CORG** framework was utilised for annotating gene regulatory elements and NCBI 35 data, and enhanced with three new features to increase annotation quality.

New methods were also developed for protein structure annotations, including a method for predicting ligand binding sites, SURFNET-ConSurf and a method for the characterisation of ligand binding sites, IsoCleft. In addition, a pipeline merging methods from several partners for the prediction of function was developed, which allows a user to send a query and runs various algorithms which provide functional annotation.

Several papers were published describing these new methods.

During the third year of the project (2006), various new methods were developed, or enhanced and published, as listed below:-

Work Package	New methods developed in 2006
1. Gene Definition/Alternative splicing	<ul style="list-style-type: none"> - EGASP – Tools and Data to validate gene structures - Tools to optimise # tissues for minimal cost but maximal coverage of transcript diversity - Prediction of Alternative Start Sites - Analysis & Experiments on heterochromatic fraction of human genome - Identifying mis-predicted proteins
2. Regulators and Promoters	<ul style="list-style-type: none"> - TRAP – Transcription Factor Affinity Prediction - Use of Multiple ChIP-chip datasets to find positive regions, especially for Transcription Start Sites
3. Expression	<ul style="list-style-type: none"> - Ranking expression experiments for human genes - Method for prediction of Enhancer Element Locator EEL
4. Variation (haplotypes and SNPs)	<ul style="list-style-type: none"> - Methods to interpret and visualise eQTLs
5. Protein families, orthologues	<ul style="list-style-type: none"> - Methods to map between domain family resources
6. Membrane Proteins and Ligands	<ul style="list-style-type: none"> - Profile-profile method to detect remote homologues for membrane proteins (SHRIMP) - Method to predict membrane protein topology and assign protein family using known motifs
7.3D protein structure	<ul style="list-style-type: none"> - Use of Pcons for model quality evaluation - Method MEPS for identifying epitopes on surface of proteins from peptide data - Method to identify appropriate models for molecular replacement in protein crystallography - Benchmark data set to test methods for choosing the best template for modelling
8. Post-translation modification and localisation	<ul style="list-style-type: none"> - Artificial neural network predictors for kinase-specific phosphorylation - SecretomeP – integrative method for protein function assignment for secreted proteins - Method for the prediction of protein features in the nucleolar proteome

9 Sequence and Structure to Function	<ul style="list-style-type: none"> - Improved tools for comparing protein binding sites - Methods to improve location and characterisation of active site (TreeDet; Xdet; desite) - Method FLORA to identify functionally distinct subfamilies - Method to identify flexible regions in proteins
10 Protein-protein complexes	<ul style="list-style-type: none"> - Combined scoring scheme for PP interactions in EciD - Improved method to derive PP interactions from protein family phylogenetic trees (mirror-trees)
11 Pathways and networks	<ul style="list-style-type: none"> - Development of Pathway Hunter Tool - Evaluation of network analysis tools for prediction and annotation

The following new methods were developed, or enhanced and published, during the fourth year:-

Work Package	New methods developed in 2007
101. Gene Definition/Alternative splicing	<ul style="list-style-type: none"> - Improved methods to identify Transcription Start Sites, integrating heterogeneous data. - Methods to survey alternative spliced variants associated with protein coding locus - Automatic pipelines for quality control of reference Human Gene Sets
102. Gene regulation and expression	<ul style="list-style-type: none"> - New methods to generate co-expressed gene clusters - New methods to identify enriched transcription factor binding in mammalian promoter regions, incorporating binding affinity predictions - Suite of matrix-based tools for localising cis-regulatory elements and modules
103. Variation (haplotypes and SNPs)	<ul style="list-style-type: none"> - Refinement of Methods to interpret eQTLs
104 Protein Function Annotation	<ul style="list-style-type: none"> - Improved methods for family clustering in Gene-3D & ProtoNet - Methods to characterise functional binding sites - Methods for function prediction - Methods to analyse flexibility and function - New automatic methods for 'Generic model quality assessment' & MODCHECK-EEKS - Protocol for mapping between family resources - G3D-BioMiner – a new pipeline for function prediction
105 Post-translation modification, membrane and localisation prediction	<ul style="list-style-type: none"> - Neural network to predict glycation sites - Lipid raft compartment predictor - Method to calculate free energy of insertion of helix into membrane, using AA contributions
106, Protein Complexes, networks and pathways	<ul style="list-style-type: none"> - Method to automatically classify enzymes based on reaction operators - Improved methods for pathway inference - Improved methods for reconstruction of metabolic models

During 2008-9, various new methods were developed or enhanced and published, as listed below:-

Work Package	New methods developed in 2008/2009
101. Gene Definition/Alternative splicing	Algorithm to detect alternative splicing associated with Alternative Transcription Start sites Improved methods to identify abnormal, incomplete and mispredicted proteins Method to identify Transcriptionally Active Regions
102. Gene regulation and expression	Algorithm to detect enriched transcription factor binding in groups of mammalian promoter regions
103. Variation (haplotypes and SNPs)	Resampling methods for Mapping QTLs
104 Protein Function Annotation	Methods to cluster protein sequences into functional clusters Method to predict subunit interactions in large complexes, using co-evolving residues FLORA – new method to discover function-related structural motifs PIGS – method for predicting immunoglobulin structures
105 Post-translation modification, membrane and localisation prediction	New method to predict acetylation New computational model for transmembrane helices in mammalian proteins New method to predict membrane topology based on free energy contributions for membrane integration efficiency

These very long lists of new/improved methods highlight the enormous productivity and excellence of the network, with major advances in almost all aspects of annotation.

(iii) Annotations Generated by the members of the consortium

In the first year of the project, we decided to use the **Human Genome 35** as a first test set and requested that all members of the consortium should contribute their annotations for these sequences. In addition to the basic data provided by Ensembl; UniProt; Interpro and MSD from the EBI – which were all made available through the BioSapiens Portal, a preliminary list of additional annotations included annotations of integral membrane proteins (IMPs), 995 mouse quantitative trait loci (QTLs) from the Mouse Genome Database; predicted catalytic residues, annotated using the Catalytic Site Atlas and PSIBLAST; membrane protein topology predictions, made using 4 different methods for all membrane proteins; predicted post-translational modifications; and domain annotations from 3 protein domain family databases.

In the second year, all members of the consortium continued to contribute their annotations to version 35 of the Human Genome and other organisms; *Caenorhabditis elegans*, *Gallus gallus*, *Mus musculus*, *Pan troglodytes*, *Rattus norvegicus*, *Tetraodon nigroviridis* and HIV-1 (strain hxb2). The number of annotations provided increased greatly and was made available through the BioSapiens Portal.

It was agreed that in order for biological inferences to be made by the biologist, a helpful categorisation of the data into groups of similar types would be greatly beneficial. With an improved order of presentation of the data on the DAS server, it would be easier to compare the same annotations from different groups, or to combine different annotations. For example, users would be able to compare the active site residues from the Catalytic Site Atlas (CSA) with those annotated within UniProt (the 'ACT_SITE' annotation) or correlate those active site residues with possible metal or nucleotide phosphate binding sites. Steps were made to utilize the 'types' command of the DAS protocol so that this could be used to retrieve a list of annotation types available for an entire data source or just for a particular segment. In the XML response of this command, each 'TYPE' tag provides a mandatory 'id' attribute, which is the unique name of the annotation type. During the second year these amendments were advertised to the network participants and the implementation of this command on the partner sites was anticipated. Once this had been implemented by all, a comprehensive list of annotations could be obtained and a categorisation by hand of all protein sequence, structure, and genomic annotations took place.

The contributions by members of the consortium in the third year have been summarised in the table below:

Work Package	Annotations made in 2006
1. Gene Definition/Alternative splicing	- Genome annotations: cow, tomato, pha aphid, many fungi & grape
2 Regulators and Promoters	- Transcription Start Sites/ Histone modification sites/Chromatic Accessibility
3 Expression	- Annotations of expression for Ensembl genomes from 38 different arrays
4 Variation (haplotypes and SNPs)	- Expression data in HS mice for hippocampi, lungs and liver
5. Protein families, orthologues	- Protein family annotations provided for protein domains in many organisms
6. Membrane Proteins and Ligands	- Annotations for human genome and for putative inner membrane proteins
8. Post-translation modification and localisation	- Catalogue of prokaryotic proteins undergoing non-classical secretion
9 Sequence and Structure to Function	- Database of structural differences for flexible proteins – STRuster web site - Improved PEDANT annotations by identification of errors - Improved functional annotations using CATH domains - Protein cognate ligands database
11 Pathways and networks	- Prediction and annotation of metabolic net (methionine salvage pathway) in a model organism - Detailed annotation of a eukaryotic signalling pathway - Metabolic model of <i>Acinetobacter baylyi</i> - Annotation of Honey Bee genome - Identification of core bacterial genome
20 Thematic WP ENCODE	- Annotations for 1% of human genome

The contributions by members of the consortium in the fourth year have been summarised in the table below:

Work Package	Annotations made in 2007
101. Gene Definition/Alternative splicing	- Experimental validation/annotation of protein translation for some ENCODE alternative spliced variants
104 Protein Function Annotation	- Improved database of orthologous protein groups in HAMAP - Multiple specific modelling projects to predict structure
105 Post-translation modification, membrane and localisation prediction	- Subcellular localisation in human genome - GPI anchors in human genome
106, Protein Complexes, networks and pathways	- Set of binary interactions for 10,000 human proteins - Functional interaction network for <i>M. pneumoniae</i> - Reconstruction of metabolic networks in 21 <i>e. coli</i> strains - Resequencing and annotation of <i>B subtilis</i> - Annotation of interactions in 184 genomes
108 Thematic WP ENCODE	- Annotations for 1% of human genome
109 Thematic WP Cancer	- Multiple Annotations on Cancer genes using both CARGO and DAS
110 Thematic WP Complex Trait Proteins	- Protein-protein interaction data was used to annotate the QTL data

All members of the consortium continued to contribute many annotations, which are summarised in the table below:

Work Package	Annotations made in 2008/2009
101. Gene Definition/Alternative splicing	Conserved Alternative Splicing between human and mouse Experimental identification of interactions between conserved non coding sequences.
102 Gene Expression	Identification of all transcription factors (TFs) and orthologous pairs in human and mouse Novel sets of cis-regulatory module predictions
103/110 Variation/Haplotypes	High resolution experimental mapping of expression QTLs in mice. Experimental Gene Expression and Copy Number Variation in mice
104 Protein Function Annotation	Prototype consensus DAS tracks
105 Post-translation modification, membrane and localisation prediction	The disulphide human proteome Experimental validation of insertion “free energy” prediction method for membrane proteins Prediction of target membrane for membrane proteins

106, Protein Complexes, networks and pathways	Functional annotation for <i>Mycoplasma pneumoniae</i> Resequencing and annotation of <i>Bacillus subtilis</i> Annotations of specific degradosome in Firmicutes and Mollicutes Annotations of Domain Interactions
109 Thematic WP Cancer	Multiple Annotations on Cancer genes using both CARGO and DAS

As above, the list of annotations has been extensive and impressive, ranging from basic gene definitions to detailed annotations of signalling networks.

Where appropriate, the annotations were made available through the BioSapiens Portal. The list of contributing servers has been combined with the DAS Registry and can be found under <http://www.dasregistry.org/>. By going to “list sources” and selecting the label “BioSapiens” a list of all BioSapiens DAS tracks will be displayed.

(iv) Thematic Work Packages

One original target of the consortium was to focus for short periods of time on specific challenges, for example a subset of proteins or genomes, of relevance for health. The first two WP were devoted to ‘HIV and HCV’ and ‘Down’s syndrome (Chromosome 21)’.

In the second year, the two thematic work packages made good progress in their efforts to coordinate annotation “horizontally” across the whole spectrum. Both work packages held their planned meetings during 2005, the infectious disease work package (WP15) in Bonn, Germany, as a combined meeting with the viRgil Network of Excellence, and the Down’s syndrome work package (WP16) in Hinxton, UK, in October 2005.

The ENCODE work package which started in 2005 had a major influence on the network, requiring close integration of the different groups and the annotations they provided. This required extensive coordination and provided interesting insights into the impact of alternative splicing on the function of proteins. Most spliced variants were found to be so radically altered that their original structure and therefore function was unlikely to be maintained. The network pursued this observation with some experiments to test if these proteins were actually expressed.

The publication from the ENCODE work package was published in 2007 in *Proceedings of the National Academy of Sciences*. Experimental validation for some of the predicted proteins, to see if they are produced in the cell, was undertaken. The ePipe tool for automatic presentation and annotation was developed ready for the scaled-up ENCODE project.

In the final reporting period, WP 108 established an automatic pipeline (EPIPE) to prepare for annotation of the complete human genome. Several new methods and tools were also developed. These will be incorporated into the pipeline in the coming year and be applied once the data from the ENCODE project for the complete human genome is available.

The Cancer work package focussed on developing the CARGO approach to presenting information about cancer genes, to complement the DAS servers. Many groups contributed annotations about the cancer genes and their mutations using this CARGO approach. In the final 18 months, WP 109 made several new tools for the annotations of potential cancer genes from many members of the network. The CARGO system which incorporates WIDGETS from many groups, allows their presentation to experimental cancer biologists. A complete analysis framework MADAS has also been developed. It was hoped to continue this work at the end of this period, in collaboration with experimental cancer groups, but unfortunately this project was not funded by the commission.

The QTL Analysis thematic WP incorporated predictions on protein-protein interactions from members of the consortium, using them to generate networks, in an attempt to identify the genes and pathways involved in generating the observed phenotype. During 2008/2009 WP110 has delivered new methods for identification of genes implicated in mouse models of human disease. By mapping QTLs and eQTLs in mouse, a useful systems biology resource for the genetic analysis of complex traits in the mouse model has been provided.

C. Spreading of Excellence Activities

Already during the first year, the Network established a permanent European School of Bioinformatics, to train bioinformaticians and to encourage best practise in the exploitation of genome annotation data for biologists. During this first period much effort was devoted to setting up a workable and durable system to coordinate the School. One basic training workshop was held in Verona, Italy, followed by the related Advanced Workshop on 'Molecular Interactions'. Another School was organised in early 2005 in Nijmegen, Netherlands. The Verona training workshop, involving lectures and practicals, was oversubscribed, with 50 participants (44% female) from 14 countries selected from 118 applications. The feedback on this course was excellent. The training material for the courses was generated and presented by the young post-doctoral scientists of the consortium, and freely available from the BioSapiens web site. The Advanced Workshop involved 25 participants from 88 applications with excellent presentations and discussions.

In addition BioSapiens provided travelling fellowships for 51 young European scientists from 18 different countries, to attend the joint ISMB/ECCB 2004 bioinformatics meeting held in Glasgow in July. This was the largest bioinformatics meeting ever held, with almost 2200 participants and for the first time joined the international ISMB and European ECCB meetings. The BioSapiens consortium organised a workshop on genome annotation at this meeting.

The two thematic work packages (which involved close interactions between the bioinformaticians and the relevant experimental community) were successful in coordinating their efforts and annotating sequence and structural data. The Down's syndrome work package held a preliminary meeting in Geneva to plan chromosome 21 re-annotation strategy. The HIV/HCV workshop was held in Germany in 2005 as a joint meeting with the viRgil NoE, which targets viral resistance.

The BioSapiens programme was presented to representatives from large pharma (19 companies) and SMEs (16 companies) in biotechnology as part of the EBI's Industry meetings. A workshop for the SMEs has been organised for September 2005.

In the second year, two more European Schools of Bioinformatics were organised, one in The Netherlands (see above) and one in Madrid, Spain, in September 2005 (followed by ECCB05). Both schools were very well attended and a long list of interested students had to be turned down.

We also consolidated the collaboration with the Bologna Winter School with the organization of a "BioSapiens corner" (<http://www.biocomp.unibo.it/~school2005/>).

Travelling fellowships were again provided for young scientists to attend the ECCB 2005 European bioinformatics meeting held in Madrid in September 2005. This was the largest bioinformatics meeting organized at the European level, with 1700 participants. The BioSapiens consortium organised the second workshop on "Genome Annotation" chaired by Rob Russell (EMBL) with more than 150 participants and 16 speakers. In the context of the conference BioSapiens also organized a workshop to discuss the structure of the collaboration between masters in bioinformatics at the European level. The meeting was attended by representatives of different masters as a follow up to the initial meeting held in Hinxton in 2004.

The collaboration with the ESF programme on "functional genomics" continued in the second year with the organisation of a session during their bi-annual conference on "functional genomics and disease" in Oslo (September 2005) in which BioSapiens sponsored a Bioinformatics session with four invited speakers, and organised a workshop to discuss the needs and views on automatic annotation with a set of invited experimental biologist.

BioSapiens was also represented during 2005 at the meetings of other large consortiums including the EuroMouse in Venice, and the ENCODE NIH (September 05 Bethesda). A consequence of this was the new thematic work package to collaborate with the ENCODE consortium.

The BioSapiens leaflets were distributed during the "Communicating European Research" meeting organized by the European Commission in Brussels in November 2005, and an article on the activities of the Network was published in the *Eur. J. Hum Genet.* (The BioSapiens Network of Excellence: a European network for integrated genome annotation. *Eur J Hum Genet.* 2005 Sep;13 (9):994-7)

Two more Schools were organised during the third project year, in Oeiras (Portugal) and in Budapest (Hungary). Both Schools were organised in collaboration with the ENFIN NoE, which contributed a post doctoral researcher to teach subjects related to Systems Biology and Modelling. The school in Portugal was organised after the meeting of one of the work packages of the ENFIN network, and the participants in the workshop were able to teach in the BioSapiens school.

Another "BioSapiens corner" was organised during the Bologna Winter School in February 2006 (<http://www.biocomp.unibo.it/~school2006/>).

As in 2004 and 2005, travelling fellowships were provided for young scientists from to attend ECCB, the European Conference on Computational Biology, held this time in Eilat (Israel). For logistic reasons the conference was held in January 2007 instead of the initially proposed date (July 2006). The BioSapiens consortium organised a third workshop on “Genome Annotation”, chaired by Nir Ben-Tal (Tel-Avid University).

As part of the activities in automatic annotation, BioSapiens was represented by Anna Tramontano at the “Automated function prediction” workshop, associated with the International Society for Computational Biology and organised by Iddo Friedberg and Adam Godzik (UCSD).

The only thematic work package in 2006 was dedicated to collaboration with ENCODE, and BioSapiens was represented at the ENCODE annual meeting (Bethesda, USA July 2006). A one day workshop was organised to assess future collaborations for the computational and experimental study of the expression and function of these proteins (BioSapiens – GENCODE workshop, EBI, Hinxton, November 2006).

In the fourth year two Schools were organised, in Valencia (Spain) and in Basel (Switzerland), in collaboration with the ENFIN NoE, which contributed to the teaching of subjects related to Systems Biology and Modelling. In the school in Valencia, I. Cases from the CNIO group introduced gene control networks and related issues, in Basel A. Di Cara introduced concepts and methods in Systems Biology.

At this point the size of our waiting list had started to decrease, and we were receiving a number of requests equivalent to the number of potential slots for each course. This probably indicates that the market for these courses was reaching saturation: the BioSapiens school has fulfilled its objectives and a new formula would be necessary in the future.

A “BioSapiens corner” was organised at the Bologna Winter School in February 2007 (<http://www.biocomp.unibo.it/~school2007/>).

For the fourth time, we provided travelling fellowships for young scientists to attend ECCB, the 6th European Conference on Computational Biology, which was co-organised with the International Society for Computational Biology (ISCB) and took place in Vienna, Austria. In connection with the conference BioSapiens co-organised the “Biosapiens_AFP Genome Annotation” Special Interest Group workshop. This year the workshop was co-chaired by Christine Orengo, representing BioSapiens.

A good part of the effort of the BioSapiens Steering committee in 2007 was devoted to the preparation of the proposal for the ESFRI programme. Finally ELIXIR was successfully organized and started with a kick-off meeting in January 2008.

The major event in 2008 was a workshop organised on April 2nd at the Université Libre de Bruxelles, Brussels, in connection with the Annual General Meeting. The workshop was entitled “From Genome to Proteome and Biological Function”, and was open to all scientists. The workshop attracted well over a hundred participants mainly from Belgium and the Netherlands.

In 2009, BioSapiens was presented at the ELIXIR Stakeholders' Meeting in Copenhagen, during an open session entitled "From information to the Medicines and Bioindustries of the Future".

The European School of Bioinformatics continued in the final period to train bioinformaticians and to encourage best practise in the exploitation of genome annotation data for biologists. During the final reporting period two schools, dedicated to basic training in bioinformatics, were organised, in Hinxton (UK) and in Brussels (Belgium). Both schools were very well attended.



Fig 1.2.2 The 8th European School of Bioinformatics at EBI, Hinxton, UK.

In total over the duration of the project, the schools trained 349 students, more than 40% of which were women. The geographical distribution of the students, the instructors and the schools was as spread out as possible. The school can be defined as a success. It created a community among the post-docs hired by the network and extended the benefits of BioSapiens to the rest of the community, not only because young scientists could participate in the school, but also because groups not included in BioSapiens were selected for running the school and appreciated this very much, and because it strengthened the collaboration between EU networks (ENFIN, EMBRACE and FELIX).

The BioSapiens Network continued their collaboration with the Bologna Winter School both in 2008 (11th – 15th February) and 2009 (2nd – 6th February).

We again provided travelling fellowships for young scientists to attend ECCB 2008 (Cagliari, Sardinia) and ECCB 2009 (Stockholm, Sweden). At ECCB 2008, BioSapiens shared a booth with EMBRACE and ENFIN, and, in collaboration with EMBRACE, organised a tutorial entitled "Interoperability of bioinformatics software and databases: how to access databases and algorithms using the Distributed Annotation System (DAS) and web-service technology".

BioSapiens also provided student and young researcher bursaries for the Biochemical Society meeting entitled “Protein Sequence, Structure and Function” which was held at the Wellcome Trust Genome Campus on 26th – 27th January 2009.

At ISMB 2008 in Toronto, Canada, BioSapiens participated in a workshop organised jointly with AFP (the Automated Function Prediction Special Interest Group).

BioSapiens participated in an Information Workshop on European Bioinformatics Resources for Small and Medium-sized Enterprises, organised jointly with ENFIN, EMBRACE and FELICS (an FP6 Integrated Infrastructures Initiative project). The workshop took place in Berlin on the 27th – 28th October, 2008, and attracted about 30 participants.

The “BioSapiens book”, correctly entitled “**Modern Genome Annotation: The BioSapiens Network**”, edited by Dmitriy Frishman and Alfonso Valencia and with contributions from most BioSapiens partners, was published in the UK by Springer on November 18th 2008 (earlier in some countries). In addition, **four review articles and a commentary were published in *Genome Biology* in early 2009:**

- Thornton, J. (for the BioSapiens Network): Annotations for all by all - the BioSapiens network *Genome Biology* 2009, **10**(2): 401.
- Juncker, AS; Jensen, LJ; Pierleoni, A; Bernsel, A; Tress, ML; Bork, P; von Heijne, G; Valencia, A; Ouzounis, CA; Casadio, R. and Brunak, S: Sequence-based feature prediction and annotation of proteins. *Genome Biology* 2009, **10**: 206.
- Loewenstein, Y; Raimondo, D; Redfern, OC; Watson, J; Frishman, D; Linial, M; Orengo, C; Thornton, J. and Tramontano, A: Protein function annotation by homology-based inference. *Genome Biology* 2009, **10**: 207.
- Harrow, J; Nagy, A; Reymond, A; Alioto, T; Patthy, L; Antonarakis, SE and Guigó, R: Identifying protein-coding genes in genomic sequences. *Genome Biology* 2009, **10**: 201.
- Vingron, M; Brazma, A; Coulson, R; van Helden, J; Manke, T; Palin, K; Sand, O. and Ukkonen, E: Integrating sequence, evolution and functional genomics in regulatory genomics. *Genome Biology* 2009, **10**(1): 202.

BioSapiens has been well represented at most major bioinformatics events of the year, through the partners, in numerous presentations. These are listed in Appendix 1 – Plan for using and disseminating knowledge: Section 2 – Dissemination of knowledge.

D. Management Actions

The consortium agreement was concluded on 17th February, 2004. The first Annual General Meeting was held in Rome in late February and approved the budget proposed by the steering committee. Further administrative procedures were agreed on, including replacement of a principal investigator, and software and test set intellectual property issues.

The Steering Committee (Thornton, Brunak, Tramontano, Valencia, Nyberg) has met regularly throughout the project, covering all aspects of the network, including finance, integration, courses and workshops. The Training Committee has met on several occasions to organise the training courses and workshops.

The second Annual General Meeting was held in Cambridge in April. The annual review was organised simultaneously, with Professor Barry Honig, Chair of the Scientific Advisory Board, as the external reviewer. No changes to the consortium were requested during 2005.

The third Annual General Meeting was held in Barcelona in March 2006. The annual review was again organised simultaneously, this time with Professor John Findlay, University of Leeds, as the external reviewer.

The following changes to the consortium were requested during 2006:

- 1) The removal of Institut Municipal d'Assistència Sanitària from, and the addition of Centre de Regulació Genòmica to, the contract, following the change of employment by one of the Principal Investigators, Roderic Guigó
- 2) The removal of Consejo Superior de Investigaciones Cientificas from, and the addition of Fundación Centro Nacional de Investigaciones Oncológicas Carlos III to, the contract, following the change of employment by one of the Principal Investigators, Alfonso Valencia
- 3) The addition of the Centre for Research and Technology Hellas to the contract, following the change of employment by one of the Principal Investigators, Christos Ouzounis (previously employed by EMBL-EBI).

The fourth Annual General Meeting (and annual review) was held at Kloster Seon, near Munich, on April 2nd – 4th. A joint BioSapiens Steering Committee/EMBRACE Executive Board meeting was held in conjunction with the ISMB/ECCB meeting in Vienna in July.

The following changes to the consortium were requested during 2007:

- The removal of Universität zu Köln from, and the addition of Technische Universität Carolo-Wilhelmina zu Braunschweig to, the contract, following the change of employment by one of the Principal Investigators, Dietmar Schomburg.

The fifth and final Annual General Meeting was held on April 2nd – 4th at the Université Libre de Bruxelles, Brussels. The annual review was organised simultaneously, with Professor John Findlay, University of Leeds, again as the external reviewer. The Steering Committee (Thornton, Brunak, Tramontano, Valencia, Nyberg) met four times in 2008.

An extension of the duration of the contract to 66 months was requested and agreed on July 4th 2008.

One change to the consortium was requested in 2008: the transfer of contractual rights and obligations from Consortium Nationale de Recherche en Génomique to Commissariat à l'Energie Atomique. This was also agreed on July 4th 2008.

1.3 Objectives Reached

The BioSapiens Network of Excellence has delivered on all 11 major objectives.

1. **Develop an integrated approach for annotation:** We have developed a fully integrated approach for annotation by implementing the DAS ‘Distributed Annotation System’ for both protein and gene annotation. Better integration of annotations was further achieved through the ‘sequence annotation ontology’. All the partners have provided DAS annotations through the BioSapiens portal (<http://www.biosapiens.info/>). As part of our thematic work package on the Cancer genes, more diverse annotations have been integrated through the CARGO infrastructure system. The ePIPE pipeline for annotating the ENCODE data on the complete human genome has also been established. We have also developed many web services, which feed into the EMBRACE catalogue for web servers.

2. **Stimulate cooperation with experimentalists:** The data used in the ENCODE and CANCER projects were generated by a large consortium of experimentalists. These co-operations have been successful, since the data are generated independently. There are also close bi-partite links between some of the computational laboratories and experimental groups around Europe. In practice, this has probably been the most difficult objective to achieve. This aspect is very ‘low-throughput’ and costly and in practice difficult to implement in a large network.

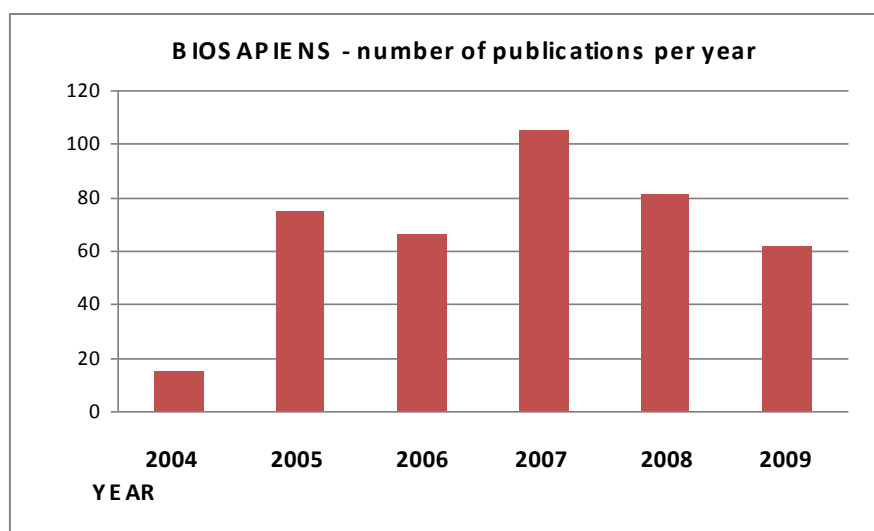


Fig 1.3.1 BioSapiens related publications per year.

3. **Develop new methods for annotation:** We have generated many new methods for computational predictions in all aspects of the network – genes, proteins and complexes and networks. This is clear from our impressive publication record (see Fig 1.3.1: over 400 publications, with over 5000 citations, are directly associated with the BioSapiens Network of Excellence), the ‘BioSapiens’ book (2008) and the series of four reviews in *Genome Biology*, written and coordinated by members of the consortium (2009).

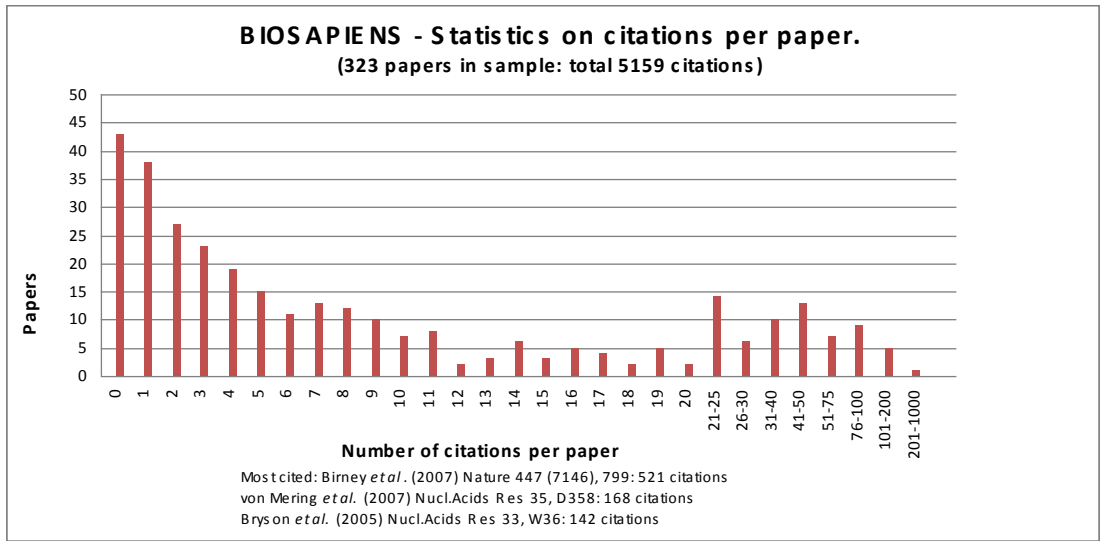


Fig 1.3.2 Citation statistics for BioSapiens related publications (note: many of the uncited papers were published in 2009).

4. **Provide Training:** We have provided extensive training through the ‘European School of Bioinformatics’ for basic bioinformatics training, which has run 9 courses, training 349 young scientists from over 20 countries throughout and outside Europe (see Fig 1.3.3). We have also provided more advanced training in numerous additional courses and workshops.

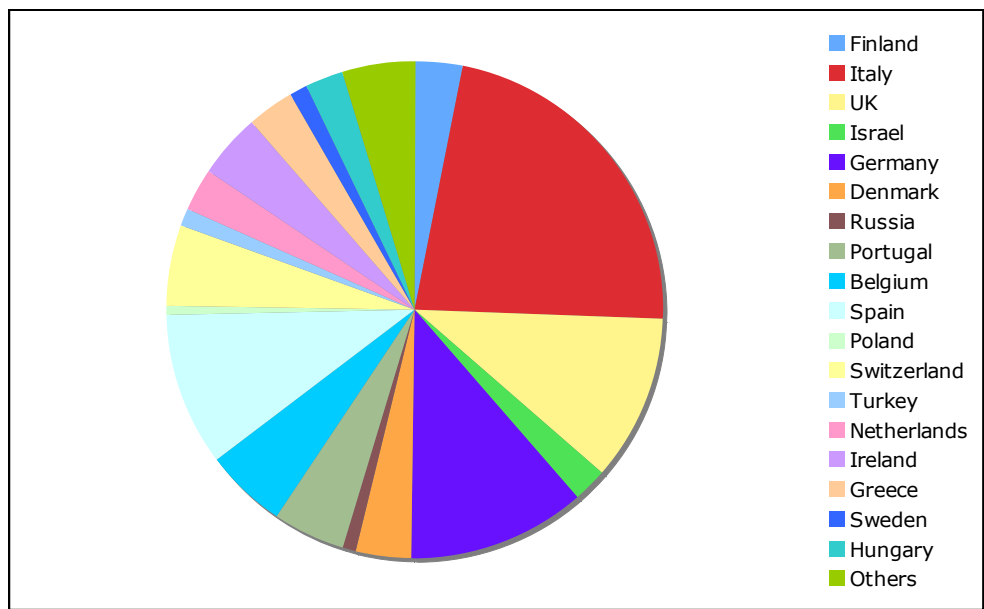


Fig 1.3.3 Geographic distribution of the 349 students trained by the Permanent School in Bioinformatics.

5. **Validate Predictions:** Some experimental validations and annotations have been performed to provide annotation of conservation and variation between genome

sequence and expression in human and mouse. In the final period, experimental annotations were provided for:

- interactions between conserved non coding sequences.
- gene expression and copy number variation in mice
- an insertion “free energy” prediction method for membrane proteins
- resequencing and annotation of *Bacillus subtilis*

6. **Provide Infrastructure for Annotation:** We have provided a full set of DAS servers and clients, which can be accessed through the DAS registry (www.DASregistry.org). With our development of the sequence annotation ontology, better integration and approaches to consensus predictions are possible. The DASTY (<http://www.ebi.ac.uk/dasty>) and SPICE clients for annotating protein features in sequences and structures respectively has been developed and extended to annotate protein-protein interactions in IntAct (see Fig 1.3.4 below). The CARGO system for presenting heterogeneous annotations in a user friendly manner, using the ‘widget’ technology has been further developed to include more widgets from other groups in the network (<http://cargo.bioinfo.cnio.es/> - see Fig 1.3.5 below). We have also developed EPipe (<http://www.cbs.dtu.dk/services/EPipe/>) – a pipeline to handle and annotate the full ENCODE data, when released (see Fig 1.3.6 below).

The screenshot displays the IntAct DASTY2 interface for the protein Inr21_human (UniProt AC: O75509). The main visualization area shows a table of annotations with columns for Feature Type, Labels, and Feature Annotations. Key annotations include:

FEATURE TYPE	LABELS	FEATURE ANNOTATIONS
disulfide crosslinked	UNIPROT: 07...	
disulfide crosslinked	UNIPROT: 07...	
glycosylated residue	UNIPROT: 07...	
polypeptide domain	Death	
polypeptide domain	THFR, DEATH 2...	
polypeptide domain	THFR, DEATH ...	
polypeptide domain	DEATH	
signal peptide	UNIPROT: 07...	
signal peptide	THFR, DEATH ...	
polypeptide repeat	THFR-Cys 1, ...	
polypeptide structural domain	DEATH like	
transmembrane	Extracellular...	
transmembrane	UNIPROT: 07...	
sufficient to bind	42..42-349....	
sufficient to bind	42..42-349....	
polypeptide conserved motif	THFR, MGRP 1	

The interface also includes a 'PROTEIN STRUCTURE' window showing a 3D ribbon diagram of the protein, and a 'MANIPULATION OPTIONS (Non positional features)' section with 'NON POSITIONAL FEATURES' and 'SEQUENCE' information.

Figure 1.3.4: DASTY2 view integrated in IntAct. The second and third features from the bottom are binding region annotations from IntAct, shown in the context of sequence annotations from UniProt and InterPro. The binding feature is shown twice because it has been shown in two independent experiments. Note the overlap of the binding feature annotation with the extracellular annotation from UniProt (two features above).



Fig 1.3.5 The CARGO web site.

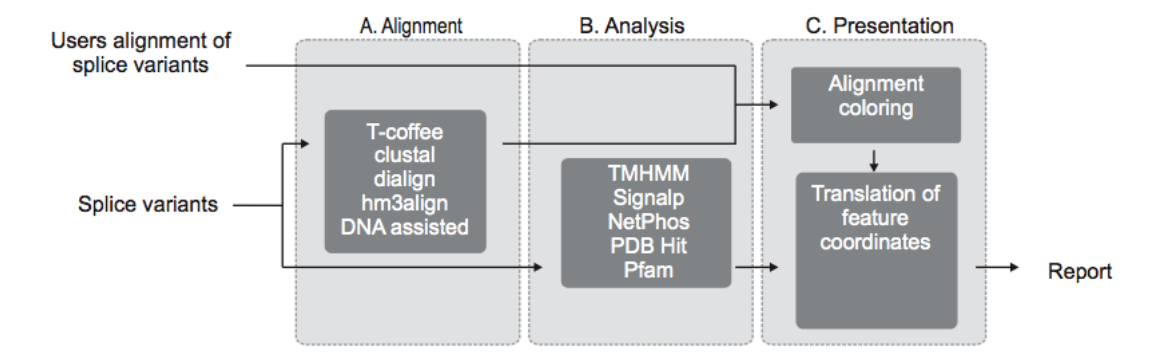


Fig. 1.3.6 General design of the automatic ENCODE pipeline (EPIPE).

7. **Provide Annotations:** We have continued to provide a large number of gene and protein annotations in all aspects of the network. These are currently available through the BioSapiens portal, and they are or will soon be visible through the global Ensembl, UniProt and IntAct data resources.

8. **Stimulate best practise:** We have promoted best practise within Europe in the use of genome knowledge captured in annotations, by organising training courses; establishing collaborations; running meetings and workshops. We have further strengthened the interactions between the research bioinformaticians around Europe and the European data service providers, mainly at EBI.

9. **Improve Training and Mobility for bioinformaticians:** We have trained and provided improved mobility of young European researchers (PhD and Post-doc) in the area of bioinformatics and annotation, by research training in individual centres and through workshops.

10. **Provide Technology transfer:** We have continuously kept the EBI's Industry Programme, involving 16 companies (which is represented by two persons on the BioSapiens Scientific Advisory Board) informed of our progress.

11. **Further a European research Area in Bioinformatics:** We have furthered a European Research area for bioinformatics, enhancing Europe's role in the academic and industrial exploitation of genomics, in several ways:

- By contributing to the organisation of the major European and international Bioinformatics conferences, and organising organised public independent meetings.
- By encouraging cooperation between laboratories spread around Europe. This is exemplified in Fig 1.3.7, which illustrates the large number of joint publications between members of the consortium.
- The BioSapiens Network has been a major influence in the development of the ELIXIR ESFRI preparatory phase to establish a sustainable infrastructure for biological information in Europe.

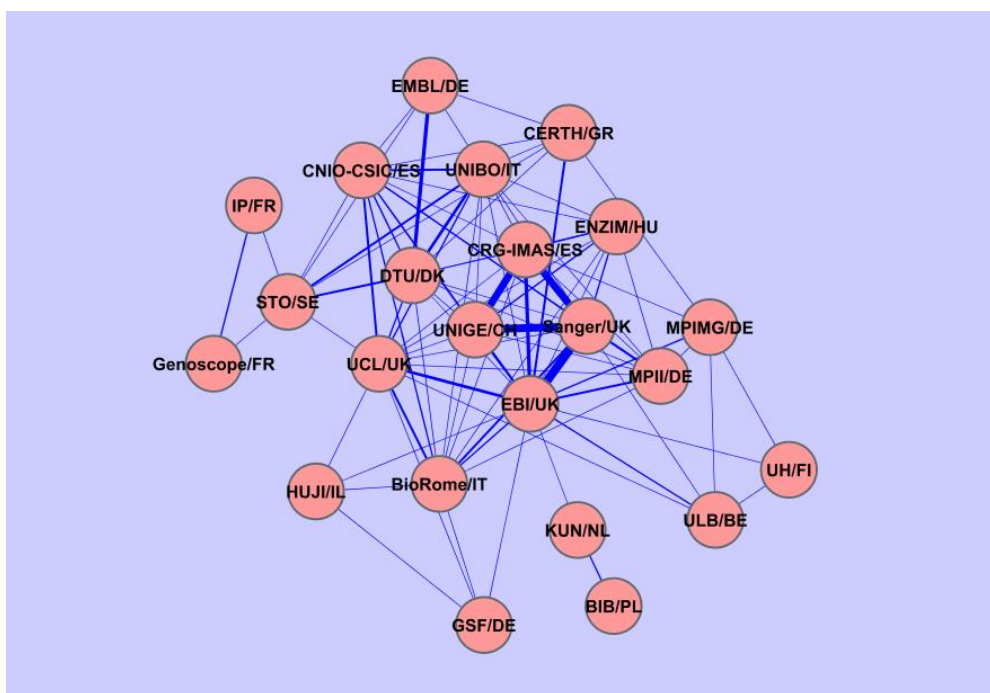


Fig. 1.3.7 Joint publications between members of the consortium.

1.4 Methodologies and approaches

In BioSapiens we have used and developed novel computational approaches both for the analysis and interpretation of biological data and the development of robust computational infrastructures, which allow better integration of gene, protein and functional annotations from laboratories throughout Europe. Building on the 'DAS' concept for distributed annotation, we have developed or improved several DAS clients for viewing data, especially for protein sequences and structures. We have also employed web services technology to integrate methods. For analysis and prediction,

many different data mining approaches have been employed in the different laboratories. To allow better integration between laboratories it proved necessary to develop the ‘protein sequence ontology’, which is now available for everyone to use.

1.5 Relationship to state-of-the-art

With the flood of sequence data now being generated by the new sequencing technologies, we clearly need to improve annotation of gene and protein sequences and structures. The major data resources have teams of curators to look after and annotate their data. The rise of Wikis and community annotation is also an important source of hand curation and should be encouraged. However, whilst hand curation will remain important for many years, it must be helped by as much automation as possible. This network has developed many new methods to improve automated annotation in particular for gene annotation, such as for promoters, annotation of alternatively spliced variants and annotation of protein function. These methods are clearly at the cutting edge of the field. The DAS registry will be maintained, allowing users and developers to view their annotations in the light of those provided by others. This will be a valuable source of annotations for both genes and proteins. The biological information infrastructure is developing very rapidly and its future shape and structure are not yet certain. It will be a mixture of centralised and distributed resources and the BioSapiens network has laid the foundation for one aspect of this complex knowledge ecosystem.

1.6 Impact of BioSapiens on industry and research sectors

The major lasting impacts of BioSapiens are:

- An infrastructure for distributed automated annotations of gene and protein sequences and structures
- New bioinformatics methods for automated annotations
- A cohort of well-trained cooperative young European bioinformaticians
- Experimentalists trained to make best use of bioinformatics tools
- A coordinated Bioinformatics Research Area in Europe
- The foundations of ELIXIR

All these impacts are relevant to life scientists of all persuasions in both the industrial and academic sectors.

To assess the impact on young scientists involved in the project, a questionnaire was sent out in early 2009 to all post doctoral researchers (or equivalent). They were asked where they were working now, what their experience of being involved in a large European project had been, and how many publications they had got out of their time with BioSapiens.

Of the 22 people who replied (out of 61), 10 were still at the same institution, but in several cases their positions had been made permanent and/or they had become independent researchers with their own funding. Four had gone to work for high-tech companies in the field of bioinformatics, and seven to other universities, including Columbia University and Yale. One post doc had left the field and become a manager

in public health care, but still felt that working in a large collaboration had helped her in her career.

All the comments made by the post docs on their experience of working within BioSapiens were positive. The comments included:

- it helped with reintegration into European science after a long period in USA
- a good opportunity to meet a great number of excellent researchers and to know about their projects
- it was very valuable as it allowed me to: 1) to extend the range of my scientific interests; 2) to start specific scientific collaboration; 3) to feel the European style of doing science; and 4) to form a general view of the European science in terms of priorities
- being involved in a collaborative project was fundamental for my career
- very enthusiastic and motivating
- valuable networking
- the possibility of cooperating with leading experts of bioinformatics was very stimulating

All respondents but two reported at least one publication, with nine publications being the record and the average three.

2. Dissemination and use

2.1 Final Plans for using and disseminating knowledge

The knowledge gained from the BioSapiens Network will continue to be made available and have an impact in the following ways:

The infrastructure developments will be maintained at the site of development. In particular

- The DASTY client will be linked to UniProt – the global resource for protein sequence and function information, and as such will become an important part of its infrastructure. It will also be incorporated into IntAct – the protein-protein interactions database at EBI.
- The DAS registry will continue to be maintained by the Wellcome Trust Sanger Institute, as part of their core programme
- The CARGO resource will be maintained at the CNIO by Alfonso Valencia's group.
- The EPipe pipeline will be maintained at the DTU in readiness to process the data released by the whole genome ENCODE project.
- Web services will be maintained by the individual partners as appropriate and entered into the EMBRACE BioCatalogue of web services, so that they can be located and accessed by all.

New methods developed in individual laboratories have mostly been published already. In the future new methods will be maintained by the relevant partner at their own web site. The annotations provided by the methods will illuminate biological research over the coming years.

Several further papers generated as part of BioSapiens are still under review and will hopefully be published in the coming year.

Knowledge generated during this project on cooperation and working practices between different European laboratories and countries will continue to be critical in the development of the ELIXIR infrastructure for biological information. This includes the training aspects of the network, which have played a major role in shaping the plans for training under ELIXIR.

2.2 Intellectual property

From the outset, the consortium agreed to a simple model, which provided an elegant mix of public access to annotation and private exploitation of the tools. All the annotations have been made available in the public domain. In contrast, the software tools developed for annotation, usually funded in part by member state support, remains the property of the individual participating laboratories to exploit as they see fit.

Appendix 1 – Plan for using and disseminating knowledge: Section 2 – Dissemination of knowledge

Planned/ actual dates	Type	Type of audience	Countries addressed	Size of audi- ence	Partner(s) responsible/ involved
January 2004	Press release	General public and scientific community	All		EBI
18 March 2004	EBI Industry Programme meeting	Professional bioinformaticians from industry	UK + other European	25	EBI
19 April 2004	SME Forum	Representatives from small and medium- sized biotechnology companies	UK + other European	15	EBI
11-12 May 2004	"Communicating European Research" Conference, Brussels http://europa.eu.int/comm/research/conferences/2004/cer2004/pdf/cer2004report.pdf	General public and scientific community	All	Approx 800	CSIC
July 2004	Project web site http://www.biosapiens.info	General public and scientific community	All		EBI
19-20 July 2004	Colloquium “experimental approach to genome annotation” organized by American Academy of Microbiology, Washington http://www.asm.org/Academy/index.asp?bid=32664	Invited scientists	USA + BioSapiens invited speaker	120	CSIC
23 June 2004	Meeting of representatives of the European masters’ courses in bioinformatics http://www.ebi.ac.uk/biosapiens/page.php?page=masters	Invited organisers of masters’ courses	EU	20	EBI BioRome CSIC

30.July 2004	Workshop on Genome Annotation (in connection with ISMB/ECCB04) http://www.ebi.ac.uk/biosapiens/page.php?page=meeting&meet=2	Scientific community	All	70	CSIC
31 July – 4 August 2004	Poster presenting BioSapiens at ISMB/ECCB04	Scientific community	All	>2000	EBI
31 July 2004	Meeting of BioSapiens representatives with Australian Bioinformatics delegation)	Interested participants	EU-Australia	15	BioRome
1 October 2004	Further meeting of BioSapiens representatives with Australian Bioinformatics delegation)	Interested participants	EU-Australia	15	BioRome
1 – 3 October 2004	Scientific meeting (EMBO sectorial meeting in bioinformatics and computational biology)	Specialist scientific community	EMBO member states		BioRome
4 October 2004	BioSapiens Scientific Advisory Board meeting	BioSapiens steering committee + SAB	All	12	BioRome
4-8 December 2004	Scientific meeting (CAPRI) http://capri.ebi.ac.uk/Gaeta.html	Specialist scientific community	All	45	BioRome, CSIC
8-10 December 2004	Scientific meeting (CASP) http://predictioncenter.llnl.gov/casp6/	Specialist scientific community	All	150	BioRome
1-4 December 2004	Presentation of BioSapiens in the “Structural Genomics & Proteomics” Joint Meeting of EU Projects” http://europa.eu.int/comm/research/press/2004/pr1911en.cfm	Specialist scientific community	All	300	DTU
11 – 12 April 2005	BioSapiens AGM	Consortium, relevant scientific community	Network partners + all	> 100 expected	EBI

22 – 26 January 2005	Second BioSapiens School in Bioinformatics, Nijmegen	Scientific community	EU	50	KUN
27 – 30 January 2005	Advanced workshop on G Protein Coupled Receptor Modelling, Nijmegen	Scientific community	EU	25	KUN
13 - 19 February 2005	Bologna Winter School	Scientific community	World wide	150	UNIBO
11-12 May 2005	BioSapiens Regulation & Gene Expression Meeting, Berlin	Scientific community	EU	20 - 30	MPIMG
11 – 12 April 2005	BioSapiens 2 nd AGM, Cambridge	The BioSapiens consortium and Scientific Advisory Board	Member countries	100	EBI
18 – 19 July 2005	“BioSapiens potential collaboration” description to the ENCODE (NIH) project	Scientist associated to ENCODE	Mainly USA	150	CSIC
6 – 10 September 2005	BioSapiens sponsored session (4 speakers) in the 2nd ESF Functional Genomics Conference , Oslo	Scientific community	World wide	150 aprox.	CSIC, EBI
8 September 2005	Biosapiens organized workshop with experimental biologist to discuss their specifics needs on the field of genome annotation in the context of 2nd ESF Functional Genomics Conference , Oslo	Scientific community	Several	20	CSIC, EBI
21 – 23 September 2005	BioSapiens-viRgil Workshop on Bioinformatics for Viral Infections, Bonn	Scientific community	EU	?	MPII
19 – 23 September 2005	Third BioSapiens School in Bioinformatics, Madrid	Scientific community	EU	40	CSIC, and Biosapiens postdocs

28 September 2005	Workshop on Genome Annotation (in connection with ECCB05) http://www.eccb05.org/workshop_session.htm	Scientific community	World wide	150	EMBL, CSIC
30 September 2005	Meeting of representatives of the European masters' courses in bioinformatics (during ECCB2005)	Scientific community	EU	30	BioRome, CSIC and others
6 – 7 October 2005	BioSapiens workshop on human chromosome 21, Hinxton	WP16 participants and collaborators	EU	30	UNIGE
14 – 15 October	EuroMouse conference, Venice	Scientific community	Mainly EU	100	EBI
14 – 15 November 2005	Communicating European Research, Brussels: distribution of BioSapiens leaflet	Scientific community and communication professionals	EU	>2000	EBI, CSIC
8 – 11 February 2006	Scientific meeting Institute for Biocomputation and Physics of Complex Systems, Zaragoza	Scientific community	Mostly Europe	100	BioRome
18 – 23 February 2006	Bologna Winter School	Scientific community	World wide	150	UNIBO
27 – 29 March 2006	BioSapiens 3 rd AGM, Barcelona	The BioSapiens consortium and Scientific Advisory Board	Member countries	100	IMAS
3 - 5 May 2006	Workshop on Sequence, Structure and Systems Approaches to Predict Protein Function, Piscataway, NJ, USA	Scientific community	World wide	200	BioRome
15 - 19 May 2006	The Fourth European School of Bioinformatics Oeiras, Portugal	Scientific community	EU	50	BioRome

1 – 3 June 2006	Italian Biochemical Society, Novara, Italy	Scientific community	Italy	200	BioRome
3 – 9 June 2006	Scientific meeting (Biological Networks) Bertinoro, Italy	Specialist scientific community	World wide	32	BioRome
24 29 June 2006	31 st FEBS Congress, ISTANBUL, Turkey	Scientific community	EU	2000	BioRome
30 August – 1 Sept 2006	Scientific meeting (Automated function prediction) University of California San Diego	Specialist scientific community	World wide	100	BioRome
4 - 8 September 2005	The Fifth European School of Bioinformatics Budapest, Hungary	Scientific community	EU	50	ENZIM
7 November 2006	BioSapiens – GENCODE workshop Hinxton, Cambridge	Scientists associated with GENCODE	Member countries, USA	30	EBI, CRG
24 November 2006	Science and society meeting (“NOW, meetings in the present continuous”) Barcelona, Spain	General	Spain	300	CSIC
26 – 30 November 2006	Scientific meeting (CASP) University of California	Specialist scientific community	World wide	150	BioRome
29 Nov – 1 Dec 2006	The RegCreative Jamboree Ghent, Belgium	Specialist scientific community	World wide	43	EBI
4 – 8 December 2006	BioInfoSummer2006, Canberra, Australia	Scientific community	World wide	100	BioRome
Jan 2007	ECCB2006, Eilat, Israel (postponed from September 2006) Workshop “Genome Annotation” Travel fellowships for young scientists	Scientific community	World wide	n/a	CSIC, EBI
21.1-21.1 2007	BioSapiens workshop, Eilat Israel	Scientific community	World wide	150	CNIO

18 – 23 February 2007	Bologna Winter School	Scientific community	World wide	150	UNIBO
26 – 28 February 2007	DAS developers workshop + DAS feature classification workshop, Hinxton	Consortium + specialist scientific community	Member countries	30	EBI, Sanger
2 – 4 April 2007	BioSapiens 4 th AGM, Kloster Seeon	The BioSapiens consortium and Scientific Advisory Board	Member countries	60	GSF
26 – 30 April 2007	6 th European School of Bioinformatics, Valencia, Spain	Scientific community	EU	40	BioRome
30 April – 2 May 2007	International Symposium on Health Informatics and Bioinformatics, Turkey	Scientific community	World wide	100	BioRome
4-6 July 2007	The 26th Leeds Annual Statistical Research Workshop	Scientific community	EU	150	BioRome
19 - 20 July 2007	BioSapiens-AFP SIG, ISMB/ECCB, Vienna, Austria	Specialist scientific community	World wide	100	UCL
27 – 31 August 2007	7 th European School of Bioinformatics, Basel, Switzerland	Scientific community	EU	40	SIB
31 Oct 2007	SME Forum meeting at EBI (jointly with ENFIN)	SMEs	EU	15	EBI, CNIO
4 – 7 Dec 2007	EMBRACE-BioSapiens workshop on web services, Manchester, UK	BioSapiens and EMBRACE	EU	15	EMBRACE NoE

11-15.2 2008 / 2-6.2.2009	Bologna Winter Schools 2008 and 2009	Scientific community	World wide	150	UNIBO
2.4.2008	BioSapiens Workshop “From Genome to Proteome and Biological Function”, Université Libre de Bruxelles	Scientific community	EU	30	EBI, ULB
2-4.4.2008	BioSapiens 5 th AGM, Brussels	BioSapiens consortium and Scientific Advisory Board	Member countries	68	EBI, ULB
22 – 24.4 2008	CASP7.5, Madrid, Spain	Scientific community	World wide	50	BioRome
12-16.5. 2008	8 th European School of Bioinformatics, Hinxton, Cambridge, UK	Scientific community	EU	40	EBI, BioRome
18-19.7 2008	Automated Function Prediction Special Interest Group meeting at ISMB2008, Toronto, Canada	Scientific community	World wide	100	HUJI
22.9.2008	“Interoperability of bioinformatics software and databases” - tutorial (with EMBRACE) at ECCB2008, Cagliari, Sardinia	Scientific community	Europe	40	EBI
22-26.9 2008	Booth distributing information on BioSapiens (jointly with EMBRACE and ENFIN) at ECCB2008, Cagliari, Sardinia	Scientific community	Europe	200	EBI
1 – 4.10 2008	3 rd ESF Conference on Functional Genomics and Disease Innsbruck, Austria	Scientific community	Europe	150 approx	EBI, CNIO, BioRome
27 - 28.10 2008	SME Forum jointly with ENFIN, EMBRACE and FELICS Berlin, Germany	Small and medium size enterprises	EU	30	EBI
20 – 21.11 2008	Presentation of BioSapiens at Networking Meeting for EU-Funded Biobanking Projects, Brussels	Scientific community/funders	EU	100 approx	CNIO
3 – 7.12 2008	CASP8, Cagliari, Italy	Scientific community	World wide	100 approx	BioRome
26 – 30.1 2009	9 th European School of Bioinformatics, Brussels, Belgium	Scientific community	EU	40	ULB, BioRome
18 – 20.5 2009	ELIXIR Stakeholders Meeting, Session “From Information to the Medicines and Bioindustries of the Future”, Copenhagen, Denmark	Scientific community, research funders	EU	150	DTU, EBI

2008	PUBLICATION: “Modern Genome Annotation: The BioSapiens Network”, edited by Dmitrij Frishman and Alfonso Valencia. Springer 2008.	Scientific community	World wide	n/a	Consortium
September 2008	Editorial in the eStrategies magazine on the EU-funded projects coordinated by the EBI (ELIXIR, EMBRACE, ENFIN and BioSapiens).	General	Europe wide	39,000	EBI, CNIO
2009	<p>PUBLICATION: BioSapiens review series in <i>Genome Biology</i></p> <ul style="list-style-type: none"> • Thornton, J. (for the BioSapiens Network): Annotations for all by all - the BioSapiens network <i>Genome Biology</i> 2009, 10(2): 401. • Juncker, AS; Jensen, LJ; Pierleoni, A; Bernsel, A; Tress, ML; Bork, P; von Heijne, G; Valencia, A; Ouzounis, CA; Casadio, R. and Brunak, S: Sequence-based feature prediction and annotation of proteins. <i>Genome Biology</i> 2009, 10: 206. • Loewenstein, Y; Raimondo, D; Redfern, OC; Watson, J; Frishman, D; Linial, M; Orengo, C; Thornton, J. and Tramontano, A: Protein function annotation by homology-based inference. <i>Genome Biology</i> 2009, 10: 207. • Harrow, J; Nagy, A; Reymond, A; Alioto, T; Patthy, L; Antonarakis, SE and Guigó, R: Identifying protein-coding genes in genomic sequences. <i>Genome Biology</i> 2009, 10: 201. • Vingron, M; Brazma, A; Coulson, R; van Helden, J; Manke, T; Palin, K; Sand, O. and Ukkonen, E: Integrating sequence, evolution and functional genomics in regulatory genomics. <i>Genome Biology</i> 2009, 10(1): 202. 	Scientific community	World wide	n/a	Consortium

Appendix 2 - List of publications

Aerts, S., van Helden, J., Sand, O. & Hassan, B. A. (2007). Fine-Tuning Enhancer Models to Predict Transcriptional Targets across Multiple Genomes. *PLoS ONE* 2, e1115. Pubmed 17973026

Alexa, A. Integrating the GO graph structure in scoring the significance of Gene Ontology terms Master Thesis, University of the Saarland, Germany (2005)

Alexa, A., Rahnenführer, J. and Lengauer, T. Improved Scoring of Functional Groups from Gene Expression Data by Decorrelating GO Graph Structure (2006) *Bioinformatics* 22 (13):1600-1607; Advance Access originally published online on April 10, 2006

Alioto, T.S. (2007) "U12DB: a database of orthologous U12-type spliceosomal introns." *Nucleic Acids Res.* 35 (DataBase Issue):D110-D115.

Amico M, Finelli M, Rossi I, Zauli A, Elofsson A, Viklund H, von Heijne G, Jones D, Krogh A, Fariselli P, Luigi Martelli P, Casadio R. (2006) PONGO: a web server for multiple predictions of all-alpha transmembrane proteins. *Nucleic Acids Res.* 34(Web Server issue):W169-72.

Andreeva, A., Prlic, A., Hubbard, T. J. & Murzin, A. G. (2007) SISYPHUS--structural alignments for proteins with non-trivial relationships. *Nucleic Acids Res* 35, D253-9.

Andres, LE; Ezkurdia, I; Garcia, B; Valencia, A; Juan, D. (2009) EclD. A database for the inference of functional interactions in E.coli. *Nucl. Acid Res.* 37, D629 – 635.

Artamonova I., Gelfand M.S. (2007) Comparative genomics and evolution of alternative splicing: the pessimists' science. *Chem Rev.* 107, 3407-3430.

Artamonova, I., Frishman, G., Frishman D. (2007) Applying negative rule mining to improve genome annotation, *BMC Bioinformatics*, 8, 261.

Ashida, H., Danchin, A. and Yokota, A. (2005) Was photosynthetic RuBisCO recruited by acquisitive evolution from RuBisCO-like proteins involved in sulfur metabolism? *Res. Microbiol.*, 156, 611-618.

Assenov, Y., Ramírez, F., Schelhorn, S.E., Lengauer, T., Albrecht, M. (2008) Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282-284. Advance Access originally published online on November 15, 2007.

Audit B, Levy ED, Gilks WR, Goldovsky L, Ouzounis CA. (2007) CORRIE: enzyme sequence annotation with confidence estimates. *BMC Bioinformatics.* 8(Suppl 4):S3.

Bahir, I., and Linial, M. (2005) ProTeus: identifying signatures in protein termini *Nucleic Acids Res.*, 33: W277 - W280.

Barbe, V., Cruveiller, S., Kunst, F., Lenoble, P., Meurice, G., Sekowska, A., Vallenet, D., Wang, T.Z., Moszer, I., Médigue, C., Danchin, A. (2009) From a consortium sequence to a unified sequence: The *Bacillus subtilis* 168 reference genome a decade later. *Microbiology* 155: 1758 – 1775.

Barta E , Sebestyén E , Pálffy TB , Tóth G , Ortutay CP and Patthy L (2005) DoOP: Databases of Orthologous Promoters, collections of clusters of orthologous upstream sequences from chordates and plants. *Nucleic Acids Res* 33 Database Issue, D86-90.

Bartoli L, Calabrese R, Fariselli P, Mita D, Casadio R. (2007) A computational approach for detecting peptidases and their specific inhibitors at the genome level- *BMC Bioinformatics* 8:Supp 1:S3.

Bartoli L, Capriotti E, Fariselli P, Martelli PL, Casadio R. (2007) The pros and cons of predicting protein contact maps. *Methods Mol Biol.* 413:199-218.

Bartoli L, Fariselli P, Casadio R (2008) The effect of backbone on the small-world properties of protein contact maps, *Phys Biol* 4:L1-L5.

Baudot, Anaïs; Real, Francisco X.; Izarzugaza, José M. G. and Valencia, Alfonso (2009) From cancer genomes to cancer models: bridging the gaps. *EMBO reports* 10, 4, 359–366.

Baudot A, Gomez G, Valencia A. (2009) Disease interpretation with molecular networks: The need for translational data integration. *Genome Biology*. In press.

Beerenwinkel, N., et al. (2005) Computational methods for the design of effective therapies against drug resistant HIV strains *Bioinformatics*, 21(21):3943-50.

Beerenwinkel, N., et al. (2005) Estimating HIV Evolutionary Pathways and the Genetic Barrier to Drug Resistance *J. Infect. Dis.*, 191(11):1953-60.

Bernsel, A., Viklund, H., Falk, J., Lindahl, E., von Heijne, G., and Elofsson, A. (2008) Prediction of membrane-protein topology from first principles. *Proc.Natl.Acad.Sci. USA* 105, 7177-7181.

Bernsel, Andreas; Viklund, Håkan; Elofsson, Arne (2008) Remote homology detection of integral membrane proteins using conserved sequence features. *Proteins: Structure, Function, and Bioinformatics*, 71, 3, 1387-1399.

Bernsel, Andreas; Viklund, Håkan; Falk, Jenny; Lindahl, Erik; von Heijne, Gunnar and Elofsson, Arne (2008) Prediction of membrane-protein topology from first principles. *Proc.Natl.Acad.Sci.* 105:7177-7181; doi:10.1073/pnas.0711151105

Bernsel, Andreas; Viklund, Håkan; Hennerdal, Aron and Elofsson, Arne (2009) TOPCONS: consensus prediction of membrane protein topology. *Nucleic Acids Res.*, Advance Access published on May 8, 2009; doi: doi:10.1093/nar/gkp363.

Bertonati, C., Tramontano, A. (2007) A model of the complex between the PfEMP1 Malaria Protein and the Human ICAM-1 Receptor. *Proteins*, 69(2), 215-222.

Birney, E. et al (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447, 7146, 799-816.

Blankenburg, H; Büch, J; Lengauer, T; Albrecht, M. (2009) DASMIweb: online integration, analysis and assessment of distributed protein interaction data. *Nucleic Acids Research. Nucl. Acids Res.* 37: W122-W128; doi:10.1093/nar/gkp438.

Blankenburg, H., Finn, R.D., Prlic, A. et al. (2009), DASMI: exchanging, annotating and assessing molecular interaction data, *Bioinformatics* 25 (10), 1321.

Blankenburg H, Ramírez F, Büch J, Albrecht M. (2009) DASMIweb: online integration, analysis and assessment of distributed protein interaction data. *Nucleic Acids Research*, 37(Web Server issue), W122-W128.

Bock, Christoph; Reither, Sabine; Mikeska, Thomas; Paulsen, Martina; Walter, Jörn and Lengauer, Thomas (2005) BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. *Bioinformatics* 21: 4067-4068; doi:10.1093/bioinformatics/bti652

Bogojeska, Jasmina; Alexa, Adrian; Altmann, André; Lengauer, Thomas and Rahnenführer, Jörg (2008) Rtreemix: an R package for estimating evolutionary pathways and genetic progression scores. *Bioinformatics* 24: 2391-2392; doi:10.1093/bioinformatics/btn410

Bojunga, Jörg; Welsch, Christoph; Antes, Iris; Albrecht, Mario; Lengauer, Thomas; Zeuzem, Stefan (2005) Structural and functional analysis of a novel mutation of CYP21B in a heterozygote carrier of 21-hydroxylase deficiency. *Hum.Genet.*, 117, 6, 558-564.

Bourguignon, PY, Danos, V, Kepes, F, Smidtas, S and Schachter V. (2006) Property-driven statistics of biological networks, in: *Trans. on Comput. Syst. Biol. VI, Transactions in Computational Systems Biology*, C. Priami and G. Plotkin (Eds.), pp. 1–15.

Brohée, S., Faust, K., Lima-Mendez, G., Sand, O., Janky, R., Vanderstocken, G., Deville, Y. & van Helden, J. (2008a). NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic Acids Res.* 36: W444-451.

Brohee, S., Faust, K., Lima-Mendez, G., Vanderstocken, G. and van

Helden, J. (2008b) Network Analysis Tools: from biological networks to clusters and pathways. *Nat Protoc* 3: 1616-1629.

Bryson, K., Cozzetto, D. & Jones, D.T. (2007) Computer-assisted protein domain boundary prediction using the DomPred server. *Curr Protein Pept Sci.*, 8: 181-188.

Bryson, K., McGuffin, L., Marsden, R., Ward, J., Sodhi, J., and Jones, D. (2005) Protein structure prediction servers at University College London *Nucleic Acids Res.*, 33:W36 - W38.

Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. *Human Mutat* (in press)

Campillos M, von Mering C, Jensen LJ, Bork P. (2006) Identification and analysis of evolutionarily cohesive functional modules in protein networks. *Genome Res.* 16(3):374-82.

Capriotti E, Arbiza L, Casadio R, Dopazo J, Dopazo H, Marti-Renom MA (2008) Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans. *Hum Mutat* 29:198-204.

Capriotti E, Calabrese R, Casadio R. (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*; 22 (22):2729-34. Epub 2006 Aug 7.

Capriotti E, Casadio R. (2007) K-Fold: a tool for the prediction of the protein folding kinetic order and rate- *Bioinformatics* 23:385-386.

Capriotti E, Fariselli P, Calabrese P, and Casadio R. (2005) Predicting protein stability changes from sequence with Support Vector Machines. *Bioinformatics* 21:ii54-ii58.

Capriotti E, Fariselli P, Casadio R (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucl. Acids Res.* 33:W303-W305.

Capriotti E, Fariselli P, Rossi I, Casadio R. (2007) A three-state prediction of single point mutations on protein stability changes- *BMC Bioinformatics* 9 Suppl 2: S6.

Capriotti, E., Fariselli, P., and Casadio, R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure *Nucleic Acids Res.*, 33:W306 - W310.

Capriotti, Emidio; Compiani, Mario (2006) Diffusion-collision of foldons elucidates the kinetic effects of point mutations and suggests control strategies of the folding process of helical proteins. *Proteins: Structure, Function, and Bioinformatics*, 64, 1, 198-209.

Carota L, Bartoli L, Fariselli P, Martelli PL, Montanucci L, Maggi G,

- Casadio R. (2008) High Throughput Comparison of Prokaryotic Genomes *Lect Notes Comp Sci* 4967:1200-1209.
- Carrabino, D., D'Onorio De Meo, P., Sanna, N., Castrignano', T. Orsini, M., Floris, M. and Tramontano, A. (2007) The mepsMAP server. *IEEE Transactions on Nanobiosciences*, 6(2), 155-161.
- Carro, Angel; Tress, Michael; de Juan, David; Pazos, Florencio; Lopez-Romero, Pedro; del Sol, Antonio; Valencia, Alfonso and Rojas, Ana M. (2006) TreeDet: a web server to explore sequence space. *Nucl. Acids Res.*34: W110-W115; doi:10.1093/nar/gkl203
- Casadio R, Helmer-Citterich M, Pesole G. (2007) Bioinformatics in Italy: BITS2006, the third annual meeting of the Italian Society of Bioinformatics- *BMC Bioinformatics* 8: Supp 1:S1.
- Casadio R, Martelli PL, Pierleoni A (2008) The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation- *Brief Funct Genomic Proteomic* 7, 63-73.
- Casadio R-Guest Editor's Introduction to the Special issue on Computational Biology and bioinformatics (WABI 2005)- Part 1. *IEE/ACM Transactions on Computational Biology and bioinformatics*, Vol.3, No4, pp321-322.
- Casadio, Rita; Fariselli, Piero; Martelli, Pier Luigi and Tasco, Gianluca (2006) Thinking the Impossible. *Protein Folding Protocols*, 2006, 305-320.
- Cases I, Pisano DG, Andres E, Carro A, Fernández JM, Gómez-López G, Rodriguez JM, Vera JF, Valencia A, Rojas AM (2007) CARGO: a web portal to integrated customised biological information. *Nucleic Acids Res* 35: W16-20.
- Castellano, S; Andrés, AM; Bosch, E; Bayes, M; Guigó, R; Clark, AG. (2009) Low exchangeability of selenocysteine, the 21st amino acid, in vertebrate proteins. *Mol Biol Evol.* MBE Advance Access published online on June 1, 2009
- Castellano, S., Lobanov, A., Chapple, C., Novoselov, S., Albrecht, M., Hua, D., Lescure, A., Lengauer, T., Krol, A., Gladyshev, V. and Guigó, R. (2005) Diversity and functional plasticity of eukaryotic selenoproteins: Identification and characterization of the SelJ family *PNAS*, 102:45, 16188-16193.
- Castellano, Sergi; Gladyshev, Vadim N.; Guigó, Roderic and Berry, Marla J. (2008) SelenoDB 1.0 : a database of selenoprotein genes, proteins and SECIS elements. *Nucl. Acids Res.* 36: D332-D338; doi:10.1093/nar/gkm731
- Castelo, R., Reymond, A., Wyss, C., Câmara, F., Parra, G., Antonarakis, S., Guigó, R. and Eyras, E. (2005) Comparative gene finding in chicken indicates that we are closing in on the set of multi-exonic widely expressed human genes *Nucleic Acids Research*, 33(6):1935-1939.

Castrignano', T., D'Onorio De Meo, P., Carrabino, D., Orsini, M., Floris, M. and Tramontano, A. (2007) The MEPS server for identifying protein conformational epitopes, *BMC Bioinformatics*, 10:1186/1471-2105-8.

Castrignano', T., D'Onorio De Meo, P., Cozzetto, D., Talamo, I. and Tramontano, A. (2006) The PMDB Protein Model Database *Nucleic Acids Research*, 34, 306-309.

Chapple, Charles E.; Guigo, Roderic; Krol, Alain (2009) SECISaln, a web based tool for the creation of structure-based alignments of eukaryotic SECIS elements. *Bioinformatics* 25, 5, 674-675.

Compiani M, Capriotti E, Casadio R (2004) Dynamics of the minimally frustrated helices determine the hierarchical folding of small helical proteins- *Phys Rev E* 69:051905.1-051905.8.

Cozzetto, D, Kryshchuk, A., Ceriani, M. and Tramontano, A. (2007) Assessment of predictions in the Model Quality Assessment category. *Proteins*, 8:175-183

Cozzetto, D., Di Matteo, A. and Tramontano, A. (2005) Ten years of prediction ... and counting *FEBS Journal*, 272, 881-882.

Cozzetto, D., Kryshchuk, A. and Tramontano, A. (2009) Evaluation of CASP8 Model Quality Predictions. *Proteins*, in press.

Cozzetto, D., Kryshchuk, A., Fidelis, K., Moulton, J., Rost, B. and Tramontano, A. (2009) Evaluation of template-based models in CASP8 with standard measures. *Proteins*, in press.

Cozzetto, Domenico and Anna Tramontano (2005) Relationship Between Multiple Sequence Alignments and Quality of Protein Comparative Models. *PROTEINS: Structure, Function, and Bioinformatics* 58:151-157.

Cozzetto, Domenico; Giorgetti, Alejandro; Raimondo, Domenico; Tramontano, Anna (2008) The Evaluation of Protein Structure Prediction Results. *Mol.Biotechnol.*, 39, 1, 1-8.

Cozzetto, Domenico; Tramontano, Anna (2005) Relationship between multiple sequence alignments and quality of protein comparative models. *Proteins: Structure, Function, and Bioinformatics*, 58, 1, 151-157.

Croes, D., Couche, F., Wodak, S. and van Helden, J. (2005) Metabolic PathFinding: inferring relevant pathways in biochemical networks *Nucleic Acids Res.*, 33, W326-W330.

Croes, D., Couche, F., Wodak, S. and van Helden, J. (2006) Inferring Meaningful Pathways in Weighted Metabolic Networks *J. Mol. Biol.*, 356, 222-236.

Danchin, A. (2009) Natural selection and immortality. *Biogerontology* [Epub ahead of print] 10.1007/s10522-008-9171-5

Danchin, A., Genome diversity: A grammar of microbial genomes. *ComPlexUs* (2004/2005) 2: 61-70

Danchin, A., (2004) The bag or the spindle: the cell factory at the time of system's biology. *Microb Cell Fact* 3: 13.

Danchin, A., A phylogenetic view of bacterial ribonucleases *Prog Nucleic Acid Res Mol Biol* (2009) 85: 1-41.

Danchin, A., Bacteria as computers making computers *FEMS Microbiol Rev* (2009) 33: 3-26

Danchin, A., Fang, G., Noria, S. (2007) The extant core bacterial proteome is an archive of the origin of life *Proteomics* 7: 875-889.

Darzentas N, Hadzidimitriou A, Murray F, Hatzi K, Josefsson P, Laoutaris N, Moreno C, Anagnostopoulos A, Jurlander J, Tsaftaris A, Chiorazzi N, Belessi C, Ghia P, Rosenquist R, Davi F, Stamatopoulos K. (2009) A different ontogenesis for chronic lymphocytic leukemia cases carrying stereotyped antigen receptors: molecular and computational evidence. *Leukemia* (in press).

de Lichtenberg, U., Jensen, T.S., Brunak, S., Bork, P. and Jensen, L.J. (2007) "Evolution of Cell Cycle Control – Same Molecular Machines, Different Regulation", *Cell Cycle*, 6, 1819–1825.

Defrance, M., Janky, R., Sand, O. and van Helden, J. (2008) Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nat Protoc.* 3: 1589-1603.

Denoed F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J, Dike S, Wyss C, Henrichsen CN, Holroyd N, Dickson MC, Taylor R, Hance Z, Foissac S, Myers RM, Rogers J, Hubbard T, Harrow J, Guigo R, Gingeras TR, Antonarakis SE, Reymond A. (2007) Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res.* 17(6):746-59.

Dessailly, B. H., Lensink, M. F., Orengo, C. A. and Wodak, S. J. (2008) LigASite--a database of biologically relevant binding sites in proteins with known apo-structures *Nucleic Acids Res* 36, D667-73. Pubmed 17933762.

Dessailly, B.H., Lensink, M.F. and Wodak, S.J. (2007). Relating destabilizing regions to known functional sites in proteins. *BMC Bioinformatics* 8, 141. Pubmed 17470296.

Dessailly, B.H., Lensink, M.F. and Wodak, S.J. (2008) LigASite: a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res* 36, D667 – 673.

Deutsch, S., Lyle, R., Dermitzakis, E., Attar, H., Subrahmanyam, L., Gehrig, C., Parand, L., Gagnebin, M., Rougemont, J., Jongeneel, C.,

Antonarakis, S. (2005) Gene expression variation and expression quantitative trait mapping of human chromosome 21 genes *Hum. Mol. Genet.*, 23:3741-9.

Dieterich, Christoph, Steffen Grossmann, Andrea Tanzer, Stefan Röpcke, Peter F. Arndt, Peter F. Stadler and Martin Vingron (2005) Comparative promoter region analysis powered by CORG. *BMC Genomics*, 6:24.

Djebali, Sarah; Kapranov, Philipp; Foissac, Sylvain; Lagarde, Julien; Reymond, Alexandre; Ucla, Catherine; Wyss, Carine; Drenkow, Jorg; Dumais, Erica; Murray, Ryan R; Lin, Chenwei; Szeto, David; Denoeud, France; Calvo, Miquel; Frankish, Adam; Harrow, Jennifer; Makrythanasis, Periklis; Vidal, Marc; Salehi-Ashtiani, Kourosh; Antonarakis, Stylianos E; Gingeras Thomas R; and Guigó, Roderic (2008) Efficient targeted transcript discovery via array-based normalization of RACE libraries. *Nat.Methods*, 5, 7, 629-635.

Domingues FS, Rahnenführer J, Lengauer T. (2007) Conformational analysis of alternative protein structures. *Bioinformatics* 23:3131-3138.

Domingues, Francisco S., Jorg Rahnenfuhrer and Thomas Lengauer (2004) Automated clustering of ensembles of alternative models in protein structure databases. *Protein Engineering, Design & Selection* vol. 17 no. 6 pp. 537–543

Duchniewicz, Marlena; Zemojtel, Tomasz; Kolanczyk, Mateusz; Grossmann, Steffen; Scheele, Jürgen S. and Zwartkruis, Fried J. T. (2006) Rap1A-Deficient T and B Cells Show Impaired Integrin-Mediated Cell Adhesion. *Mol. Cell. Biol.* 26: 643-653

Durot M, Bourguignon PY, Schachter V. (2009) Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol Rev.* 33(1):164-90.

Durot M, Le Fèvre F, de Berardinis V, Kreimeyer A, Vallenet D, Combe C, Smidtas S, Salanoubat M, Weissenbach J, Schachter V. (2008) Iterative reconstruction of a global metabolic model of *Acinetobacter baylyi* ADP1 using high-throughput growth phenotype and gene essentiality data. *BMC Syst Biol.* 2:85.

Dutilh, Bas E.; He, Ying; Hekkelman, Maarten L. and Huynen, Martijn A. (2008) Signature, a web server for taxonomic characterization of sequence samples using signature genes. *Nucl. Acids Res.* 36: W470-W474; doi:10.1093/nar/gkn277

Eitner, K., Gaweda, T., Hoffmann, M., Jura, M., Rychlewski, L., Barciszewski, J., (2007) eHiTS-to-VMD Interface Application. The Search for Tyrosine-tRNA Ligase Inhibitors. *J Chem Inf Model.* 47 (2):695-702.

Enquist, K., Fransson, M., Boekel, C., Bengtsson, I., Geiger, K., Lang, L., Pettersson, A., Johansson, S., von Heijne, G., and Nilsson, I.M. (2009) Membrane-integration characteristics of two ABC transporters, CFTR and P-glycoprotein. *J.Mol.Biol.* 387, 1153-1164.

Ezkurdia I, Bartoli L, Fariselli P, Casadio R, Valencia A, Tress ML. (2009) Progress and challenges in predicting protein-protein interaction sites. *Brief Bioinform* 10:233-246.

Fang, G, Rocha, EPC, Danchin, A. (2005) How essential are non-essential genes? *Mol Biol Evol* 22: 2147-2156.

Fang, G., Ho, C., Qiu, YW, Cubas, V, Yu, Z, Cabau, C, Cheung, F, Moszer, I, Danchin, A. (2005) Specialized microbial databases for inductive exploration of microbial genome sequences. *BMC Genomics* 6:14.

Fang, G., Rocha, E.P.C., Danchin, A. (2008) Persistence drives gene clustering in bacterial genomes. *BMC Genomics*, 9, 4.

Fang, Gang, Christine Ho, Yaowu Qiu, Virginie Cubas, Zhou Yu, Cedric Cabau, Frankie Cheung, Ivan Moszer, Antoine Danchin (2005) Specialized microbial databases for inductive exploration of microbial genome sequences, *BMC Genomics*, 6:14

Fariselli P, Finelli M, Rossi I, Amico M, Zauli A, Martelli PL, Casadio R (2005) TRAMPLE: the transmembrane protein labelling environment. *Nucl. Acids Res.* Jul 1;33(Web Server issue):W198-201.

Fariselli P, Martelli PL, Casadio R (2005) The posterior-Viterbi: a new decoding algorithm for hidden Markov models. *BMC Bioinformatics* 6 Suppl 4:S12

Fariselli P, Molinini D, Casadio R, Krogh A (2007) Prediction of structurally-determined coiled-coil domains with Hidden Markov Models- (Hochreiter S, Wagner R, editors) *Bioinformatics research and development*, Springer ed. *Lect Notes Comp Sci*, LNBI 4414:292-302.

Fariselli P, Rossi I, Capriotti E, Casadio R (2007) The WWWH of remote homolog detection: the state of the art. *Brief Bioinform* 8:78-87.

Faust, K., Croes, D. and van Helden, J. (2009). Metabolic Pathfinding Using RPAIR Annotation. *J Mol Biol.* 388: 390-414.

Freilich S, Goldovsky L, Ouzounis CA, Thornton JM. (2008) Metabolic innovations towards the human lineage. *BMC Evol Biol.* 2008 ;8247.

Frishman, D. (2007) Protein annotation at genomic scale: the current status. *Chemical Reviews*, 107 (8), 3448-3466.

Furney SJ, Higgins DG, Ouzounis CA, López-Bigas N., (2006) Structural and functional properties of genes involved in human cancer. *BMC Genomics* 7, 3.

Gasbarra, D, Pirinen, M, Kulathinal, S. and Sillanpää, MJ (2009) Estimating haplotype frequencies by combining data from large DNA pools with database information. *IEEE/ACM Transactions on*

Gauthier, NP; Erup Larsen, M; Wernersson,R; de Lichtenberg, U; Juhl Jensen, L; Brunak, S; Skøt Jensen, T. (2008) Cyclebase.org—a comprehensive multi-organism online database of cell-cycle experiments. *Nucleic Acids Research*, Vol. 36, Database issue D854-D859.

Geiger, K., Hedin, L., Bernsel, A., Hennerdahl, A., Illergård, K., Enquist, K., Kauko, A., Cristobal, S., von Heijne, G., Lerch-bader, M., Nilsson, I.M., and elofsson, A. (2009) Membrane insertion of marginally hydrophobic transmembrane helices depends on sequence context. Submitted.

George, Richard A.; Spriggs, Ruth V.; Bartlett, Gail J.; Gutteridge, Alex; MacArthur, Malcolm W.; Porter, Craig T.; Al-Lazikani, Bissan; Thornton, Janet M. and Swindells, Mark B. (2005) Inaugural Article: Effective function annotation through catalytic residue conservation. *Proc.Natl.Acad.Sci.* 102; 12299-12304; doi:10.1073/pnas.0504833102

Gianazza, E., Vergani, L., Brizio, C., Brambilla, D., Begum, S., Giancaspero, T.A., Conserva, F., Eberini, I., Angelini, C., Pegoraro, E., Tramontano, A. and Barile, M. (2006) Coordinated and reversible reduction of enzymes involved in oxidative metabolism in skeletal muscle mitochondria from a riboflavin-responsive, multiple acyl-CoA dehydrogenase deficiency (RR-MAD) patient *Electrophoresis*, 27 1182-1198.

Gianni T, Martelli PL, Casadio R, Campadelli-Fiume G (2005) An internal fusion peptide in herpes simplex virus glycoprotein H enables fusion required for infection of cells- *J Virol* 79:2931-2940.

Giorgetti, A; Raimondo, D; Miele, AE; Tramontano, A. (2005) Evaluating the usefulness of protein structure models for molecular replacement. *Bioinformatics*, 21 Suppl 2:ii72-6.

Glaser, Fabian; Morris, Richard J.; Najmanovich, Rafael J.; Laskowski, Roman A.; Thornton, Janet M. (2006) A method for localizing ligand binding pockets in protein structures. *Proteins: Structure, Function, and Bioinformatics*, 62, 2, 479-488.

Godard, P., Urrestarazu, A., Vissers, S., Kontos, K., Bontempi, G., van Helden, J. & Andre, B. (2007). Effect of 21 different nitrogen sources on global gene expression in the yeast *Saccharomyces cerevisiae*. *Mol Cell Biol* 27, 3065-86. Pubmed 17308034.

Gómez-López, Gonzalo and Valencia, Alfonso (2008) Bioinformatics and cancer research: building bridges for translational research. *Clinical and Translational Oncology*, 10, 2, 85-95.

Graña, O., Baker, D., MacCallum, R., Meiler, J., Punta, M., Rost, B., Tress, M. and Valencia, A. (2005) CASP6 assessment of contact prediction *Proteins: Structure, Function, and Bioinformatics*, 61, 214-

224.

Graña, O., Eyrich, V., Pazos, F., Rost, B., and Valencia, A. (2005) EVAcon: a protein contact prediction evaluation service *Nucleic Acids Res.*, 33:W347 - W351.

Graña, Osvaldo; Baker, David; MacCallum, Robert M.; Meiler, Jens; Punta, Marco; Rost, Burkhard; Tress, Michael L.; Valencia, Alfonso (2005) CASP6 assessment of contact prediction. *Proteins: Structure, Function, and Bioinformatics*, 61, S7, 214-224.

Grandi F, Sandal M, Guarguaglini G, Capriotti E, Casadio R, Samori B. (2006) Hierarchical mechanochemical switches in angiostatin. *Chembiochem.* (11):1774-82.

Grano V, Tasco G, Casadio R, Diano N, Portaccio M, Rossi S, Bencivenga U, Compiani M, De Maio A, Mita DG (2004) Reduction of active elastase concentration by means of immobilized inhibitors: a novel therapeutic approach- *Biotechnol Prog* 20:968-974.

Granseth, E; Daley, DO; Rapp, M; Melén, K; von Heijne, G. (2005) Experimentally Constrained Topology Models for 51,208 Bacterial Inner Membrane Proteins. *J Mol Biol* 352, 489-494.

Greene, Lesley H.; Lewis, Tony E.; Addou, Sarah; Cuff, Alison; Dallman, Tim; Dibley, Mark; Redfern, Oliver; Pearl, Frances; Nambudiry, Rekha; Reid, Adam; Sillitoe, Ian; Yeats, Corin; Thornton, Janet M. and Orengo, Christine A. (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucl. Acids Res.*35: D291-D297; doi:10.1093/nar/gkl959

Griffith, OL; Montgomery, SB; Bernier, B; Chu, B; Kasaian, K; Aerts, S; Mahony, S; Sleumer, MC; Bilenky, M; Haeussler, M; Griffith, M; Gallo, SM; Giardine, B; Hooghe, B; Van Loo, P; Blanco, E; Ticoll, A; Lithwick, S; Portales-Casamar, E; Donaldson, IJ; Robertson, G; Wadelius, C; De Bleser, P; Vlieghe, D; Halfon, MS; Wasserman, W; Hardison, R; Bergman, CM; Jones, CJM; and The Open Regulatory Annotation Consortium (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Research*, Vol. 36, Database issue D107-D113.

Ginalski, K; Grishin, NV; Godzik, A; Rychlewski, L. (2005) Practical lessons from protein structure prediction. *Nucleic Acids Res.* 2005 ;33 (6):1874-91.

Guariento, M., Raimondo, D., Assfalg, M., Zanzoni, S., Pesente, P., Ragona, L., Tramontano, A. and Molinari, H. (2008) Identification and functional characterization of the bile acid transport proteins in non-mammalian ileum and mammalian liver. *Proteins*, 70(2):462-72.

Guigó, R, Flicek, P, Abril, JF, Reymond, A, Lagarde, J, Denoeud, F, Antonarakis, S, Ashburner, M, Bajic, VB, Birney, E, Castelo, R, Eyras, E, Ucla, C, Gingeras, T, Harrow, J, Hubbard, T, Lewis, S and Reese MG.

(2006) EGASP: The human ENCODE Genome Annotation Assessment Project. *Genome Biology*, 7(Suppl 1):S2.

Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J. (2006) Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* 124(1):47-59.

Harrington, Eoghan D.; Jensen, Lars J.; Bork, Peer (2008) Predicting biological networks from genomic data *FEBS Lett.*, 582, 8, 1251-1258.

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. (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*;7 Suppl 1:S4.1-9.

Harrow J, Nagy A, Reymond A, Alioto T, Patthy L, Antonarakis SE, Guigó R. (2009) Identifying protein-coding genes in genomic sequences. *Genome Biol.* 10, 201.

Hekkelman, M., and Vriend, G. (2005) MRS: a fast and compact retrieval system for biological data *Nucleic Acids Res.*, 33:W766 - W769.

Hessa, T., Meindl-Beinker, N.M., Bernsel, A., Kim, H., Sato, Y., Lerch-Bader, M., Nilsson, IM., White, S.H., and von Heijne, G. (2007) Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature* 450, 1026-1030.

Hinsby AM, Kiemer L, Karlberg EO, Lage K, Fausboll A, Juncker AS, Andersen JS, Mann M, Brunak S, (2006) A Wiring of the Human Nucleolus, *Molecular Cell*, 22, 285-295.

Hoffmann, Robert; Krallinger, Martin; Andres, Eduardo; Tamames, Javier; Blaschke, Christian and Valencia, Alfonso (2005) Text Mining for Metabolic Pathways, Signaling Cascades, and Protein Networks. *Sci. STKE* 10 May 2005: pe21.

Holliday, GL; Mitchell, JBO and Thornton, JM (2009) Understanding the Functional Roles of Amino Acid Residues in Enzyme Catalysis *J. Mol. Biol.* (2009) 390, 560–577.

Huang, GJ; Shifman, S; Valdar, W; Johannesson, M; Yalcin, B; Taylor, MS; Taylor, JM; Mott, R; Flint, J. (2009) High resolution mapping of expression QTLs in heterogeneous stock mice in multiple tissues. *Genome Res.* 19(6):1133-40.

Hubbard, T.J., Aken, B.L., Ayling, S. et al. (2009), Ensembl 2009, *Nucleic Acids Res* 37 (Database issue), D690.

Inberg, Alex, Michal Linial (2004) Evolutional insights on uncharacterized SARS coronavirus genes. *FEBS Letters* 577, 159–164.

Izarzugaza JMG, Baresic A, McMillan L EM, Yeats C, Clegg AB,

Orengo CA, Martin ACR and Valencia A (2009) An integrated approach to the interpretation of Single Amino Acid Polymorphisms within the framework of CATH and Gene3D BMC Bioinform. In the press

Izarzugaza, JM; Juan, D; Pons, C; Pazos, F; Valencia, A (2008) Enhancing the prediction of protein pairings between interacting families using orthology. BMC Bioinformatics 9, 35.

Izarzugaza JM, Juan D, Pons C, Ranea JA, Valencia A, Pazos F. (2006) TSEMA: interactive prediction of protein pairings between interacting families. Nucleic Acids Research. 34, W315-9.

Izarzugaza JM. G., Redfern O. C., Orengo C. A., and Valencia A. (2009) Cancer associated mutations are preferentially distributed in protein kinase functional sites. Proteins in the press

Janky, R. & Helden, J. (2007). Discovery of conserved motifs in promoters of orthologous genes in prokaryotes. Methods Mol Biol 395, 293-308.

Janky, R. and van Helden, J. Evaluation of phylogenetic footprint discovery for the prediction of bacterial cis-regulatory elements (2008). BMC Bioinformatics 9:37.

Jenkinson AM, Albrecht M, Birney E, Blankenburg H, Down T, Finn RD, Hermjakob H, Hubbard TJ, Jimenez RC, Jones P, Kähäri A, Kulesha E, Macías JR, Reeves GA, Prlic A. (2008) Integrating biological data - the Distributed Annotation System, BMC Bioinformatics 9 Suppl 8, S3.

Jensen LJ, Jensen TS, de Lichtenberg U, Brunak S, Bork P. (2006) Co-evolution of transcriptional and post-translational cell-cycle regulation. Nature. 5;443(7111):594 –7.

Jensen, L.J, Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks T. and Bork, P. (2008) “eggNOG: automated construction and annotation of orthologous groups of genes”, Nucleic Acids Research, 36, D250-254.

Jensen, Lars J.; Kuhn, Michael; Stark, Manuel; Chaffron, Samuel; Creevey, Chris; Muller, Jean; Doerks, Tobias; Julien, Philippe; Roth, Alexander; Simonovic, Milan; Bork, Peer and von Mering Christian (2009) STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucl. Acids Res. 37: D412-D416; doi:10.1093/nar/gkn760.

Jimenez, Rafael C.; Quinn, Antony F.; Garcia, Alexander; Labarga, Alberto; O'Neill, Kieran; Martinez, Fernando; Salazar, Gustavo A.; and Hermjakob, Henning (2008) Dasty2, an Ajax protein DAS client Bioinformatics 24: 2119-2121; doi:10.1093/bioinformatics/btn387

Johannesson, Martina; Lopez-Aumatell, Regina; Stridh, Pernilla; Diez, Margarita; Tuncel, Jonatan; Blazquez, Gloria; Martinez-Membrives, Esther; Canete, Toni; Vicens-Costa, Elia; Graham, Delyth; Copley, Richard R.; Hernandez-Pliego, Polinka; Beyeen, Amennai D.;

Ockinger, Johan; Fernandez-Santamaria, Cristina; Gulko, Percio S.; Brenner, Max; Tobena, Adolf; Guitart-Masip, Marc; Gimenez-Llort, Lydia; Dominiczak, Anna; Holmdahl, Rikard; Gauguier, Dominique; Olsson, Tomas; Mott, Richard; Valdar, William; Redei, Eva E.; Fernandez-Teruel, Alberto; Flint, Jonathan (2009) A resource for the simultaneous high-resolution mapping of multiple quantitative trait loci in rats: The NIH heterogeneous stock. *Genome Res.* 19: 150-158; doi:10.1101/gr.081497.108

Johansen, Morten Bo; Kiemer, Lars and Brunak, Søren (2006) Analysis and prediction of mammalian protein glycation. *Glycobiology* 16: 844-853; doi:10.1093/glycob/cwl009

Jones, D. T.; Bryson, K.; Coleman, A.; McGuffin, L.J.; Sadowski, M.I.; Sodhi, J.S.; Ward, J.J. (2005) Prediction of novel and analogous folds using fragment assembly and fold recognition. *Proteins: Structure, Function, and Bioinformatics*, 61, S7, 143-152.

Jones, D.T. (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. *Bioinformatics*. 23: 538-544.

Jones P, Vinod N, Down T, Hackmann A, Kähäri A, Kretschmann E, Quinn A, Wieser D, Hermjakob H, Apweiler R. (2005) Dasty and UniProt DAS: a perfect pair for protein feature visualization *Bioinformatics* 21, 3198-9.

Juan, David; Pazos, Florencio and Valencia, Alfonso (2008) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proc.Natl.Acad.Sci.* 105: 934-939; doi:10.1073/pnas.0709671105

Juan, David; Pazos, Florencio; Valencia Alfonso (2008) Co-evolution and co-adaptation in protein networks *FEBS Lett.*, 582, 8, 1225-1230.

Juncker AS, Jensen LJ, Pierleoni A, Bernsel A, Tress ML, Bork P, von Heijne G, Valencia A, Ouzounis CA, Casadio R, Brunak S. (2009) Sequence-based feature prediction and annotation of proteins. *Genome Biol* 10:206.

Kahraman, A., Morris, R.J., Laskowski, R.A., Thornton, J.M. (2007) Shape Variation, Protein Binding Pockets and their Ligands. *JMB.* 368: 283 – 301.

Kaján, L., Rychlewski, L. (2007) Evaluation of 3D-Jury on CASP7 models. *BMC Bioinformatics.* 8: 304.

Kamburov A, Goldovsky L, Freilich S, Kapazoglou A, Kunin V, Enright AJ, Tsafaris A, Ouzounis CA (2007) Denoising inferred functional association networks obtained by gene fusion analysis. *BMC Genomics*; 8, 8460.

Kaplan, N. and Linial, M. (2007) ProtoBee: Classification and global annotation scheme for the honey bee genome. *Genome Research* 16,

1431-1439.

Kaplan, N., Sasson, O., Inbar, U., Friedlich, M., Fromer, M., Fleischer, H., Portugaly, E., Linial, N., and Linial, M. (2005) ProtoNet 4.0: A hierarchical classification of one million protein sequences *Nucleic Acids Res.*, 33:D216 - D218.

Kaplan, N; Morpurgo, N; Linial, M (2007) Novel Families of Toxin-like Peptides in Insects and Mammals: A Computational Approach. *J. Mol.Biol.* 369, 553-556.

Kaplan, Noam, Moriah Friedlich, Menachem Fromer and Michal Linial (2004) A functional hierarchical organization of the protein sequence space. *BMC Bioinformatics*, 5:196

Kärkkäinen, J. and Na J. C. (2007) Faster Filters for Approximate String Matching. *Proc. Algorithm Engineering and Experiments (ALENEX07), SIAM 2007.*

Kauko, A., Hedin, L., Thebaud, E., Cristobal, S., Elofsson, A., and von Heijne, G. (2009). Repositioning of transmembrane alpha-helices during membrane protein folding. Submitted.

Ketter, Ralf; Kim, Yoo-Jin; Storck, Simone; Rahnenführer, Jörg; Romeike, Bernd F. M.; Steudel, Wolf-Ingo; Zang, Klaus D.; Henn, Wolfram (2007) Hyperdiploidy defines a distinct cytogenetic entity of meningiomas *J.Neurooncol.*, 83, 2, 213-221

Kiemer, L., Bendtsen, J.D., and Blom, N. (2005) NetAcet: Prediction of N-terminal acetylation sites *Bioinformatics*, 21(7):1269-70.

Kinch, LN; Ginalski, K; Rychlewski, L; Grishin, NV (2005) Identification of novel restriction endonuclease-like fold families among hypothetical proteins. *Nucleic Acids Res.* 33 (11):3598-605.

Knizewski, L, Kinch, L, Grishin, NV, Rychlewski, L, Ginalski, K. (2006) Human Herpesvirus 1 UL24 Gene Encodes a Potential PD-(D/E)XK Endonuclease. *J Virol.* 80: p 2575-7.

Knizewski, L., Kinch, L., Grishin, N., Rychlewski, L., Ginalski, K. (2007) Realm of PD-(D/E)XK nuclease superfamily revisited: detection of novel families with modified transitive meta profile searches. *BMC Struct Biol.* 7 (1):40.

Koivisto, M, Rastas, P, Mannila, H. and Ukkonen, E (2008) Phasing genotypes using a hidden Markov model. In: I. Mandiou & A. Zelikovsky (eds.), *Bioinformatics Algorithms: Techniques and Applications*, pp. 355-372, Wiley Book Series on Bioinformatics, Wiley.

Kokoszynska, K; Ostrowski, J; Rychlewski, L; Wyrwicz, L. (2008) The fold recognition of CP2 transcription factors gives new insights into the function and evolution of tumor suppressor protein p53. *Cell Cycle.* 2008 Sep 15;7(18):2907-15.

- Korhonen, J, Martinmaki, P, Pizzi, C, Rastas, P and Ukkonen, E (2009) MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics* (in review).
- Koussounadis, A., Redfern, O.C. & Jones, D.T. (2009) Improving classification in protein structure databases using text mining. *BMC Bioinformatics*. 10:129.
- Krallinger et al. (2009) Extraction of human kinase mutations from literature, databases and genotyping studies, *BMC Bioinformatics SI mutations – in the press*.
- Lage, K; Karlberg, EO; Størling, M; Ólason, PÍ; Pedersen, AG; Rigina, O; Hinsby, AM; Tümer, Z; Pociot, F; Tommerup, N; Moreau, Y and Brunak, S. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat.Biotechnol.*, 25, 3, 309-316.
- Laskowski, R., Watson, J., and Thornton, J. (2005) ProFunc: a server for predicting protein function from 3D structure *Nucleic Acids Res.*, 33:W89 - W93.
- Laskowski, Roman A.; Watson, James D.; Thornton, Janet M. (2005) Protein Function Prediction Using Local 3D Templates. *J.Mol.Biol.*, 351, 3, 614-626.
- Le Fèvre F, Smidtas S, Combe C, Durot M, d'Alché-Buc F, Schachter V. (2009) CycSim--an online tool for exploring and experimenting with genome-scale metabolic models. *Bioinformatics*. 25(15):1987-8.
- Le Fèvre F., Smidtas S., Schächter V. (2007) Cyclone: Java-based querying and computing with Pathway Genome Databases, *Bioinformatics*, 15; 23 (10):1299-300.
- Lee, D., A. Grant, R. Marsden and C. Orengo (2005) Identification and Distribution of Protein Families in 120 Completed Genomes using Gene3D. *Proteins*. 59:603–615.
- Lehmann, Clara; Däumer, Martin; Boussaad, Ibrahim; Sing, Tobias; Beerenwinkel, Niko; Lengauer, Thomas; Schmeisser, Norbert; Wyen, Christoph; Fätkenheuer, Gerd; Kaiser, Rolf (2006) Stable co-receptor usage of HIV in patients with ongoing treatment failure on HAART. *Journal of Clinical Virology*, 37, 4, 300-304.
- Lensink MF, Mendez R., Wodak SJ. (2007) Docking and Scoring Protein Complexes: CAPRI 3rd Edition. *Proteins* 69(4):704-18. Pubmed 17918726
- Lensink, M.F. and Mendez, R. (2008) Recognition-induced conformational changes in protein-protein docking. *Curr. Pharmaceut. Biotech.* 9, 77-75, Pubmed 18393864.
- Leon, Eduardo Andres; Ezkurdia, Iakes; García, Beatriz; Valencia, Alfonso and Juan, David (2009) EcID. A database for the inference of

functional interactions in *E. coli*. *Nucl. Acids Res.* 37: D629-D635; doi:10.1093/nar/gkn853.

Lima-Mendez, G., van Helden, J., Toussaint, A. and Leplae, R. (2008). Reticulate representation of evolutionary and functional relationships between phage genomes. *Molecular Biology and Evolution* 25:762-777.

Linding, R., Jensen, L.J, Pasculescu, A., Olhovsky, M., Colwill, K., Bork, P., Yaffe, M.B. and Pawson, T. (2008) "NetworKIN: a resource for exploring cellular phosphorylation networks", *Nucleic Acids Research*, 36, D695-699.

Linding, R., Jensen, L.J., Ostheimer, G.J., van Vugt, M.A.T.M., Jørgensen, C., Miron, I.M., Diella, F., Colwill, K., Taylor, L., Elder, K., Metalnikov, P., Nguyen, V., Pasculescu, A., Jin, J., Park, J.G., Samson, L.D., Woodgett, J.R., Russell, R.B., Bork, P., Yaffe, M.B., and Pawson, T. (2007) "Systematic discovery of in vivo phosphorylation networks", *Cell*, 129, 1415–1426.

Lobley, A., Swindells, M.B., Orengo, C.A. & Jones, D.T. (2007) Inferring function using patterns of native disorder in proteins. *PLoS Comput. Biol.* 3:e162.

Lobley, A.E.; Nugent, T; Orengo, C.A. and Jones, D.T. (2008) FFPred: an integrated feature-based function prediction server for vertebrate proteomes *Nucl. Acids Res.* 36: W297-W302; doi:10.1093/nar/gkn193

Lobley, A; Sadowski, M.I. & Jones, D.T. (2009) pGenTHREADER and pDomTHREADER: New Methods For Improved Protein Fold Recognition and Superfamily Discrimination. *Bioinformatics.* 25, 1761-1767.

Loewenstein, Yaniv and Linial, Michal (2008) Connect the dots: exposing hidden protein family connections from the entire sequence tree. *Bioinformatics* 24: i193-i199; doi:10.1093/bioinformatics/btn301

Loewenstein, Yaniv; Portugaly, Elon; Fromer, Menachem and Linial, Michal (2008) Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space. *Bioinformatics* 24: i41-i49; doi:10.1093/bioinformatics/btn174

Loewenstein, Y., Raimondo, D., Redfern, O., Watson, J. Frishman, D., Linial, M., Orengo, C., Thornton, J., Tramontano, A. (2009) Protein function annotation by homology-based inference *Genome Biology*, 10:207.

López G, Rojas A, Tress M, Valencia A. (2007) Assessment of predictions submitted for the CASP7 function prediction category. *Proteins.* 69 Suppl 8:165-74.

López G, Valencia A, Tress M. (2007) FireDB--a database of functionally important residues from proteins of known structure. *Nucleic Acids Res.* 35(Database issue):D219-23.

López G, Valencia A, Tress ML. (2007) firestar--prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res.* 35(Web Server issue):W573-7.

López-Bigas N, Blencowe BJ, Ouzounis CA (2006) Highly consistent patterns for inherited human diseases at the molecular level. *Bioinformatics* 22, 269-277.

Lukk, M; Kapushesky, M; Nikkila, J; Parkinson, H; Goncalves, A; Huber, W; Ukkonen, E; Brazma, A.: A global map of major transcriptional states of the human genome. *Nature Biotechnology* (submitted).

Lyle, R., Prandini, P., Osoegawa, K., ten Hallers, B., Humphray, S., Zhu, B., Eyraes, E., Castelo, R., Bird, C., Gagos, S., Scott, C., Cox, A., Deutsch, S., Ucla, C., Cruets, M., Dahoun, S., She, X., Bena, F., Wang, S.Y, Van Broeckhoven, C., Eichler, E.E., Guigo, R., Rogers, J., de Jong, P., Reymond, A. & Antonarakis, S.E. (2007) Islands of euchromatin-like sequence and expressed sequences within heterochromatic regions of the human genome: initial sequence analysis of 21p, *Genome Res.* 17: 1690 – 1696.

Makita, Y., de Hoon, M.J.L., Danchin, A. (2007). Hon-yaku: a biology-driven Bayesian methodology for identifying translation initiation sites in prokaryotes *BMC Bioinformatics* 8: 47.

Manke, T., Roider, H. and Vingron, M. (2008) Statistical Modeling of Transcription Factor Binding Affinities Predicts Regulatory Interactions. *PLOS Comp Biol.* 4(3): e1000039.

Marani P, Wagner S, Baars L, Genevoux P, de Gier JW, Nilsson I, Casadio R, von Heijne G. (2006) New *Escherichia coli* outer membrane proteins identified through prediction and experimental verification. *Protein Sci.* (4):884-9. Epub 2006 Mar 7.

Marcatili, P., Bussotti, G., Tramontano, A. (2007) The MoVin server. *BMC Bioinformatics*, 9 Suppl 2:S11.

Marcatili, P., Rosi, A. and Tramontano, A. PIGS: Automatic prediction of antibody structures (2008) *Bioinformatics*, 24(17):1953-1954.

Marsden, R., Lee, D., Maibaum, M., Yeats, C., and Orengo, C. (2006) Comprehensive genome analysis of 203 genomes provides structural genomics with new insights into protein family space *Nucleic Acids Res.*, 34 (3). pp. 1066-1080.

Marsden, Russell L; Ranea, Juan A.G; Sillero, Antonio; Redfern, Oliver; Yeats, Corin; Maibaum, Michael; Lee, David; Addou, Sarah; Reeves, Gabrielle A; Dallman, Timothy J; and Orengo, Christine A (2006) Exploiting protein structure data to explore the evolution of protein function and biological complexity. *Phil. Trans. R. Soc. B* 361:425-440; doi:10.1098/rstb.2005.1801

Masojc, B., Medrek, K., Debniak, T., Lubinski, J., Wyrwicz, L., Koczyk,

G., Hoffmann, M., Rychlewski, L. (2007) ARLTS1 Trp149Stop Mutation and the Risk of Ovarian Cancer. *Cancer Res.* 67 (9):4533.

Mayr G, Domingues FS, Lackner P. (2007) Comparative analysis of protein structure alignments. *BMC Struct Biol* 7:50.

McGuffin, L. J., Smith R. T., Bryson, K., Sorensen, S. A. & Jones, D. T. (2006) High throughput profile-profile based fold recognition for the entire Human proteome. *BMC Bioinformatics*, 7, 288.

Mechold, U, Ogryzko, V, Ngo, S, Danchin, A. (2006) Oligoribonuclease is a common downstream target of lithium-induced pAp accumulation in *Escherichia coli* and human cells. *Nucleic Acids Res* 34: 2364-2373.

Mechold, U., Fang, G., Ngo, S., Ogryzko, V., Danchin, A. (2007) : YtqI from *Bacillus subtilis* has both oligoribonuclease and pAp-phosphatase activity *Nucleic Acids Res* 35: 4552-4561.

Médigue, C, Krin, E, Pascal, G, Barbe, V, Bernsel, A, Bertin, N, Cheung, F, Cruveiller, S, D'Amico, S, Duilio, A, Fang, G, Feller, G, Ho, C, Mangenot, S, Marino, G, Nilsson, J, Parrilli, E, Rocha, EPC, Rouy, Z, Sekowska, A, Tutino, ML, Vallenet, D, von Heijne, G, Danchin, A. (2005) Coping with cold: the genome of the versatile marine Antarctica bacterium *Pseudoalteromonas haloplanktis* TAC125. *Genome Res* 15: 1325-1335.

Méndez, Raúl; Leplae, Raphaël; Lensink, Marc F.; Wodak, Shoshana J. (2005) Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins: Structure, Function, and Bioinformatics*, 60, 2, 150-169.

Mihm, Ulrike; Ackermann, Oliver; Welsch, Christoph; Herrmann, Eva; Hofmann, Wolf Peter; Grigorian, Natalia; Welker, Martin Walter; Lengauer, Thomas; Zeuzem, Stefan; Sarrazin, Christoph (2009) Clinical relevance of the 2'-5'-oligoadenylate synthetase/RNase L system for treatment response in chronic hepatitis. *C J.Hepatol.*, 50, 1, 49-58.

Miller, Martin Lee; Jensen, Lars Juhl; Diella, Francesca; Jørgensen, Claus; Tinti, Michele; Li, Lei; Hsiung, Marilyn; Parker, Sirlester A.; Bordeaux, Jennifer; Sicheritz-Ponten, Thomas; Olhovsky, Marina; Pasculescu, Adrian; Alexander, Jes; Knapp, Stefan; Blom, Nikolaj; Bork, Peer; Li, Shawn; Cesareni, Gianni; Pawson, Tony; Turk, Benjamin E.; Yaffe, Michael B.; Brunak, Søren and Linding, Rune (2008) Linear Motif Atlas for Phosphorylation-Dependent Signalling. *Sci. Signal.* 1 (35), ra2. [DOI: 10.1126/scisignal.1159433]

Montanucci L, Fariselli P, Martelli PL, Casadio R . (2008) Predicting protein thermostability changes from sequence upon multiple mutations. *Bioinformatics* 24:i190-i195.

Montanucci L, Fariselli P, Martelli PL, Rossi I, Casadio R (2008) In Silico Evidence of the Relationship Between miRNAs and siRNAs. *Open Appl Inf Journal* 2:9-13.

Montanucci L, Martelli PL, Fariselli P, Casadio R. (2007) Robust determinants of thermostability highlighted by a codon frequency index capable of discriminating thermophilic from mesophilic genomes- *J Proteome Res* 6:2502-8.

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B. and Tramontano, A. (2009) Critical Assessment of Protein Structure Prediction (CASP) - Round VIII. *Proteins*, in press.

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Hubbard, T. and Tramontano, A (2007). Critical Assessment of Protein Structure Prediction (CASP) - Round VII. *Proteins*, 8:3-9.

Moult, J., Fidelis, K., Rost, B., and Hubbard, T. and Tramontano, A. (2005) Critical Assessment of Methods of Protein Structure Prediction (CASP) - Round VI, *Proteins*, S7, 3-7.

Murray F, Darzentas N, Hadzidimitriou A, Tobin G, Boudjogra M, Scielzo C, Laoutaris N, Karlsson K, Baran-Marzsak F, Tsaftaris A, Moreno C, Anagnostopoulos A, Caligaris-Cappio F, Vaur D, Ouzounis C, Belessi C, Ghia P, Davi F, Rosenquist R, Stamatopoulos K. (2008) Stereotyped patterns of somatic hypermutation in subsets of patients with chronic lymphocytic leukemia: implications for the role of antigen selection in leukemogenesis. *Blood*. 111(3):1524-1533.

Naamane, N., van Helden, J. & Eizirik, D. L. (2007). In silico identification of NF-kappaB-regulated genes in pancreatic beta-cells. *BMC Bioinformatics* 8, 55.

Nagy, A., Hédi Hegyi, Krisztina Farkas, Hedvig Tordai, Evelin Kozma, László Bányai, László Patthy (2008) Identification and correction of abnormal, incomplete and mispredicted proteins in public databases. *BMC Bioinformatics*. 9, 353.

Nugent, T. & Jones, D.T. (2009) Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*. 10:159.

Nugent, Timothy; Mole, Sara E.; Jones, David T. (2008) The transmembrane topology of Batten disease protein CLN3 determined by consensus computational prediction constrained by experimental data *FEBS Lett.*, 582, 7, 1019-1024.

O'Neill, K; Garcia, A; Schwegmann, A; Jimenez, RC; Jacobson, D; Hermjakob, H. (2008) OntoDas – a tool for facilitating the construction of complex queries to the Gene Ontology. *BMC Bioinformatics*, 9:437.

Occhino, M., Ghiotto, F., Soro, S., Mortarino, M., Bosi, S., Maffei, M. Bruno, S., Nardini, M., Figini, M., Tramontano, A., Ciccone, E. (2008) Dissecting the structural determinants of the interaction between the HCMV UL18 protein and the CD85j immune receptor. *J. Immunol.*, 15;180(2):957-68.

Occhipinti E, Bec N, Gambirasio B, Baietta G, Martelli PL, Casadio R, Balny C, Lange R, Tortora P. (2006) Pressure and temperature as tools

for investigating the role of individual non-covalent interactions in enzymatic reactions *Sulfolobus solfataricus* carboxypeptidase as a model enzyme. *Biochim Biophys Acta*. 1764(3):563-72. Epub 2006 Jan 10.

Olason PI. (2005) Integrating protein annotation resources through the Distributed Annotation System, *Nucleic Acids Res.*, 33, 468-470.

Oliva, R., Cavallo, L. and Tramontano, A. (2006) Accurate energies of hydrogen bonded nucleic acid base pairs and triplets in tRNA tertiary interactions, *Nucl. Acids Res.* 34(3), 865-879.

Ott MA, Vriend G (2006) Correcting ligands, metabolites, and pathways. *BMC Bioinformatics.*;7: 517.

Pagel, P., Oesterheld, M., Tovstukhina, O., Strack, N., Stuempflen, V., Frishman, D. (2008) DIMA 2.0: \hat{A} -Predicted and Known Domain Interactions. *Nucl. Acids Res*, 36 (Database issue):D651-D655.

Paiva, AC; Oliveira, L; Horn, F; Bywater, RP; Vriend, G. (2006) Modeling GPCRs. *Ernst Schering Found Symp Proc*, (2):23-47.

Palin, Kimmo (2007) Computational Methods for Locating and Analyzing Conserved Gene Regulatory DNA Elements, Report A-2007-7 (Doctoral dissertation), University of Helsinki, Department of Computer Science.

Palin K., Taipale J., Ukkonen E. (2006) Locating potential enhancer elements by comparative genomics using the EEL software. *Nature Protocols* 1(1), 368-374, 27.

Parkinson, Helen; Kapushesky, Misha; Kolesnikov, Nikolay; Rustici, Gabriella; Shojatalab, Mohammad; Abeygunawardena, Niran; Berube, Hugo; Dylag, Mirosław; Emam, Ibrahim; Farne, Anna; Holloway, Ele; Lukk, Margus; Malone, James; Mani, Roby; Pilicheva, Ekaterina; Rayner, Tim F.; Rezwan, Faisal; Sharma, Anjan; Williams, Eleanor; Bradley, Xiangqun Zheng; Adamusiak, Tomasz; Brandizi, Marco; Burdett, Tony; Coulson, Richard; Krestyaninova, Maria; Kurnosov, Pavel; Maguire, Eamonn; Neogi, Sudeshna; Guha, Rocca-Serra; Philippe, Sansone; Susanna-Assunta, Sklyar; Nataliya, Zhao; Mengyao, Sarkans; Ugis and Brazma, Alvis (2009) ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucl. Acids Res.* 2009 37: D868-D872; doi:10.1093/nar/gkn889

Parra, G, Reymond, A, Dabbouseh, N, Dermitzakis, ET, Castelo, R, Thompson, TM, Antonarakis, SE and Guigó, R. (2006) *Genome Res.* 16: 37-44; doi:10.1101/gr.4145906

Pas J, Wyszko E, Rolle K, Rychlewski L, Nowak S, Zukiel R, Barciszewski J. (2006) Analysis of structure and function of tenascin-C. *Int J Biochem Cell Biol.* 38(9), p.1594-602.

Pascal, G, Médigue, C, Danchin, A. (2005) Universal biases in protein composition of model prokaryotes. *Proteins* 60: 27-35.

Pascal, G, Médigue, C, Danchin. (2006) Persistent biases in the amino acid composition of prokaryotic proteins. *Bioessays* 28: 726-738.

Patthy, L. (2006) Evolution of multidomain proteins. In: *Evolutionary Genetics. Concepts and Case Studies*. Edited by Charles W. Fox and Jason B. Wolf. Oxford University Press, pp. 211-221.

Patthy, L. (2007) The Evolution of Proteome Complexity and Diversity. In: *Evolutionary Genomics and Proteomics*. Edited by Mark Page and Andrew Pomiankowski. (Sinauer Associates).

Pazos, Florencio and Valencia, Alfonso (2008) Protein co-evolution, co-adaptation and interactions. *EMBO J.* 27(20): 2648–2655.

Pazos, Florencio; Rausell, Antonio and Valencia Alfonso (2006) Phylogeny-independent detection of functional residues. *Bioinformatics* 22: 1440-1448; doi:10.1093/bioinformatics/btl104

Pellegrini-Calace, M. and Tramontano, A. (2006) Identification of a novel putative mitogen-activated kinase cascade on human chromosome 21 by computational approaches. *Bioinformatics*, 22, 775-778.

Penzkofer T, Dandekar T, Zemojtel T. (2005) L1Base: from functional annotation to prediction of active LINE-1 elements. *Nucleic Acids Res.* 33, Database Issue: D498-500.

Pettitt, Chris S.; McGuffin, Liam J.; and Jones, David T. (2005) Improving sequence-based fold recognition by using 3D model quality assessment. *Bioinformatics* 21: 3509-3515; doi:10.1093/bioinformatics/bti540

Pierleoni A, Martelli PL, Casadio R (2009) PredGPI: a GPI anchor predictor. *BMC Bioinformatics* 10:233.

Pierleoni A, Martelli PL, Fariselli P, Casadio R. (2006) Related Articles, Links BaCellLo: a balanced subcellular localization predictor. *Bioinformatics* 22(14):e408-16.

Pierleoni A, Martelli PL, Fariselli P, Casadio R. (2007) BaCellLo: a Balanced subCellular Localization predictor- *Nature Protocols* DOI:10.1038/nprot.2007.165.

Pierleoni A, Martelli PL, Fariselli P, Casadio R. (2007) eSLDB: eukaryotic subcellular localization database. *Nucleic Acids Res* 35:D208-12.

Pirinen, M, Kulathinal, S, Gasbarra, D and Sillanpää, MJ (2008) Estimating population haplotype frequencies from pooled DNA samples using PHASE algorithm. *Genetics Research* 90, 509-524.

Plewczynski D, von Grotthuss M, Spieser SA, Rychlewski L, Wyrwicz LS, Ginalski K, Koch U. (2007) Target Specific Compound Identification using Support Vector Machine. *Comb Chem High*

Throughput Screen. 10(3):189-96.

Plewczynski, D, Hoffmann, M, von Grothuss, M, Ginalski, K, Rychlewski, L. (2007) In Silico prediction of SARS protease inhibitors by virtual high throughput screening. *Chem Biol Drug Des.* 69(4):269-79.

Plewczynski, D, Spieser, S, Koch, U. (2006) Assessing Different Classification Methods for Virtual Screening *J. Chem. Inf. Model.* 46(3), p.1098-106.

Plewczynski, D, Tkacz, A, Wyrwicz, L, Godzik, A, Kloczkowski, A and Rychlewski, L. (2006) Support Vector Machine Classification of Linear Functional Motifs in Proteins *Journal of Molecular Modelling* 12(4), p. 453-61.

Plewczyński, D., Hoffmann, M., Knizewski, L., Rychlewski, L., Eitner, K., Ginalski, K. (2007) Modelling of potentially promising SARS protease inhibitors. *Journal of Physics: Condensed Matter* 28; 285207.

Plewczynski, D., Hoffmann, M., von Grothuss, M., Ginalski, K., Rychlewski, L. (2007) In Silico Prediction of SARS Protease Inhibitors by Virtual High Throughput Screening. *Chem Biol Drug Des.* 69 (4):269-79.

Plewczynski, D; Hoffmann, M; von Grothuss, M; Knizewski, L.; Rychlewski, L; Eitner, K; Ginalski, K. (2007) Modelling of Potentially Promising SARS Protease Inhibitors. *J. Phys.: Condens. Matter* 19, 285207.

Plewczynski, D; Jaroszewski, L; Godzik, A; Kloczkowski, A; Rychlewski, L (2005) Molecular modeling of phosphorylation sites in proteins using a database of local structure segments. *J Mol Model.* 11 (6):431-8.

Plewczyński, D., Slabinski, L., Tkacz, A., Kajan, L., Holm, L., Ginalski, K., Rychlewski, L. (2007) The RPSP: Web server for prediction of signal peptides. *Polymer* 19; 5493-5496.

Plewczynski, D., Tkacz, A., Wyrwicz, L., Rychlewski, L., Ginalski, K. (2008) AutoMotif Server for prediction of phosphorylation sites in proteins using support vector machine: 2007 update. *J Mol Model.* 14 (1):69-76.

Plewczynski, D., von Grothuss, M., Spieser, S.A.H., Rychlewski, L., Wyrwicz, L.S., Ginalski, K., Koch, U., (2007) Target specific compound identification using a support vector machine. *Comb Chem High Throughput Screen.* 10 (3):189-96.

Plewczynski, D; von Grothuss, M; Rychlewski, L; Ginalski, K. (2009) Virtual high throughput screening using combined random forest and flexible docking. *Comb Chem High Throughput Screen.* 12(5):484-9.

Plewczynski, Dariusz (2009) kNNsim: k-Nearest neighbors similarity with genetic algorithm features optimization enhances the prediction of

activity classes for small molecules. *Journal of Molecular Modeling*, 15, 6, 591-596.

Plewczynski, D; Ginalski, K. (2009) The interactome: predicting the protein-protein interactions in cells. *Cell Mol Biol Lett*. 14(1):1-22.

Plewczynski, D; Rychlewski, L. (2009) Meta-basic estimates the size of druggable human genome. *Journal of Molecular Modeling*, 15, 6, 695-699.

Plewczynski, D; Rychlewski, L; Ye, Y; Jaroszewski, L; Godzik, A. (2004) Integrated web service for improving alignment quality based on segments comparison. *BMC Bioinformatics*. 22; 5:98.

Plewczynski, D; Slabinski, L; Ginalski, K; Rychlewski, L (2008) Prediction of signal peptides in protein sequences by neural networks. *Acta Biochim Pol*. 55(2):261-7.

Plewczynski, D; Tkacz, A; Wyrwicz, LS and Rychlewski, L (2005) AutoMotif server: prediction of single residue post-translational modifications in proteins. *Bioinformatics* 2005 21: 2525-2527; doi:10.1093/bioinformatics/bti333

Plewczyński, D; Tkacz, A; Godzik, A; Rychlewski, L. (2005) A support vector machine approach to the identification of phosphorylation sites. *Cell Mol Biol Lett*. 10 (1):73-89.

Portugaly E, Harel A, Linial N and Linial M. (2006) EVEREST: Automatic identification and classification of protein domains in all protein sequences. *BMC Bioinformatics*. 7, 277

Portugaly E, Linial N and Linial M. (2007) EVEREST: A collection of evolutionary conserved protein domains. *Nucleic Acids Res*. 35, D241-246.

Prandini, Paola; Deutsch, Samuel; Lyle, Robert; Gagnebin, Maryline; Delucinge Vivier, Celine; Delorenzi, Mauro; Gehrig, Corinne; Descombes, Patrick; Sherman, Stephanie; Bricarelli, Franca; Dagna Baldo, Chiara; Novelli, Antonio; Dallapiccola, Bruno; Antonarakis, Stylianos E. (2007) Natural Gene-Expression Variation in Down Syndrome Modulates the Outcome of Gene-Dosage Imbalance. *Am.J.Hum.Genetics*, 81, 2, 252-263.

Prlić A, Down TA, Kulesha E, Finn RD, Kähäri A, Hubbard TJ. (2007) Integrating sequence and structural biology with DAS. *BMC Bioinformatics*. 8:333.

Rahmenführer, Jörg; Beerenwinkel, Niko; Schulz, Wolfgang A.; Hartmann, Christian; von Deimling, Andreas; Wullich, a Bernd and Lengauer, Thomas (2005) Estimating cancer survival and clinical outcome based on genetic tumor progression scores. *Bioinformatics* 21: 2438-2446; doi:10.1093/bioinformatics/bti312

Raimondo, D., Giorgetti, A., Bosi, S. and Tramontano, A. (2007) An

automatic procedure for using models of proteins in molecular replacement. *Proteins* 15; 66(3):689-96.

Raimondo, D., Giorgetti, A., Miele, AE., Tramontano, A. (2005) Evaluating the usefulness of protein structure models for molecular replacement *Bioinformatics*, 21(S2), 72-76.

Raimondo, D., Giorgetti, A., Bernassola, F., Melino G., Tramontano, A. (2008) Modelling and molecular dynamics of the interaction between the E3 ubiquitin ligase Itch and the E2 UbcH7. *Molecular Pharmacology*, 76 (11): 1620-1627.

Ralser, M; Nonhoff, U; Albrecht, M; Lengauer, T; Wanker, EE; Lehrach, H; Krobitsch, S (2005) Ataxin-2 and huntingtin interact with endophilin-A complexes to function in plastin-associated pathways. *Human Molecular Genetics*, doi:10.1093/hmg/ddi321.

Ramírez, F., Schlicker, A., Assenov, Y., Lengauer, T., Albrecht, M. (2007) Computational analysis of human protein interaction networks. *Proteomics* 7:2541-2552.

Ranea JG, Yeats C, Grant A, Orengo CA (2007) Predicting Protein Function with Hierarchical Phylogenetic Profiles: The Gene3D Phylo-Tuner Method Applied to Eukaryotic Genomes *PLOS Comp Biol* 3(11): e237.

Ranea, J., Yeats, C., Marsden, R., and Orengo, C. (2007) Gene3D and Understanding Proteome Evolution; in *Structural approaches to sequence evolution: Molecules, networks, populations*. Ch: 2 (35-54) Editors: Bastolla, U., Porto, M., Roman, H.E. and Vendruscolo, M.) Springer-Verlag.

Rapp, Mikaela; Granseth, Erik; Seppälä, Susanna and von Heijne, Gunnar (2006) Identification and evolution of dual-topology membrane proteins. *Nat.Struct.Mol.Biol.*, 13, 2, 112-116.

Rastas, P. (2009) A General Framework for Local Pairwise Alignment Statistics with Gaps. *WABI 2009* (in press).

Rastas, P., Koivisto, M., Mannila, H. and Ukkonen, E. (2008): Phasing genotypes using a hidden Markov model. In: I. Mandoiu and A. Zelikovsky (eds.), *Bioinformatics Algorithms: Techniques and Applications*, pp. 373-391, Wiley.

Rastas, P, Kollin, J. and Koivisto, M. (2008) Fast Bayesian Haplotype Inference Via Context Tree Weighting. *WABI 2008*: 259-270

Rastas, P. and Ukkonen, E. (2007) Haplotype Inference Via Hierarchical Genotype Parsing. *WABI 2007*: 85-97. *Lecture Notes in Computer Science* 4645, Springer.

Reeves, G.A., Eilbeck, K., Magrane, M. et al. (2008), The Protein Feature Ontology: A Tool for the Unification of Protein Feature Annotations, *Bioinformatics* 24 (23), 2767.

Reeves, GA; Thornton JM; BioSapiens Network of Excellence. (2006) Integrating biological data through the genome. *Hum Mol Genet.* 15 Spec No 1:R81-7. Review.

Reeves, Gabrielle A.; Talavera, David; Thornton, Janet M. (2009) Genome and proteome annotation: organization, interpretation and integration. *J R Soc Interface* 6, 31, 129-147.

Reid, Adam James; Yeats, Corin and Orengo, Christine Anne (2007) Methods of remote homology detection can be combined to increase coverage by 10% in the midnight zone. *Bioinformatics* 23: 2353-2360; doi:10.1093/bioinformatics/btm355

Riley, L., Schmidt, T., Wagner, C., Volz, A., Artamonova, I., Heumann, K., Mewes, H.W., Frishman, D. (2007) PEDANT genome database: ten years online. *Nucl. Acids. Res.*, 35 (Database issue). D354-7.

Roider HG, Manke T, O'Keefe S, Vingron M and Haas SA. (2009) PASTAA: identifying transcription factors associated with sets of co-regulated genes. *Bioinformatics* 25(4):435-42.

Roider, H., Kanhere, A., Manke, T. and Vingron, M (2007) Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics* 23 (2): 134--41.

Rychlewski, Leszek (2005) Molecular modeling of phosphorylation sites in proteins using a database of local structure segments. *Journal of Molecular Modeling*, 11, 6, 431-438.

Sadowski, M. I.; Jones, D. T. (2007) Benchmarking template selection and model quality assessment for high-resolution comparative modelling. *Proteins: Structure, Function, and Bioinformatics*, 69, 3, 476-485.

Sadowski, M.I. & Jones, D.T. (2009) An automatic method for assessing structural importance of amino acid positions. *BMC Struct Biol.* 9,10.

Sand, O. & Helden, J. (2007). Discovery of motifs in promoters of coregulated genes. *Methods Mol Biol* 395, 329-48.

Sand, O., Thomas-Chollier, M. and van Helden, J. (2009) Retrieve-ensembl-seq: user-friendly and large-scale retrieval of single or multi-genome sequences from Ensembl. Submitted on May 25th 2009.

Sand, O., Thomas-Chollier, M., Turatsinze, J.-V., Janky, R., Defrance, M., Vervisch, E., Brohée, S., van Helden, J. (2008a). RSAT: Regulatory Sequence Analysis Tools. *Nucleic Acid Res.* 36: W119-127.

Sand, O., Thomas-Chollier, M., Vervisch, E. and van Helden, J. (2008b) Analyzing multiple data sets by interconnecting RSAT programs via SOAP Web services-an example with ChIP-chip data. *Nat Protoc.* 3: 1604-1615.

Sander O, Sing T, Sommer I, Low AJ, Cheung PK, Harrigan PR, Lengauer T, Domingues FS. (2007) Structural Descriptors of gp120 V3 Loop for the Prediction of HIV-1 Coreceptor Usage. *PLoS Comput Biol* 3:e58.

Sander O, Sommer I, Lengauer T. (2006) Local protein structure prediction using discriminative models *BMC Bioinformatics*, 7:14.
Saric J, Jensen LJ, Ouzounova R, Rojas I, Bork P. (2006) Extraction of regulatory gene/protein networks from Medline. *Bioinformatics*. 22(6):645-50.

Sarrazin, C., Mihm, U., Herrmann, E., Welsch, C., Albrecht, M., Sarrazin, U., Traver, S., Lengauer, T., Zeuzem, S. et al. (2005) Clinical significance of in vitro replication-enhancing mutations of the hepatitis C virus (HCV) replicon in patients with chronic HCV infection. *J. Infect Dis.*, 192(10):1710-9.

Schächter V. (2007) *Heterogeneous Molecular Networks in Kepes F. (Ed), Biological Networks, World Scientific, ISBN 978-981-270-695-9.*
Schelhorn, Sven-Eric; Lengauer, Thomas and Albrecht Mario (2008) An integrative approach for predicting interactions of protein regions. *Bioinformatics* 24: i35-i41; doi:10.1093/bioinformatics/btn290

Schlicker A, Albrecht M. (2008) FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Res*, 36: D434-D439; doi:10.1093/nar/gkm806.

Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. (2006) A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics* 7:302 (2006).

Schlicker A, Rahnenführer J, Albrecht M, Lengauer T, Domingues FS. (2007) GOTax: investigating biological processes and biochemical activities along the taxonomic tree. *Genome Biol.* 8:R33

Schlicker, A., Huthmacher, C., Ramírez, F., Lengauer, T., Albrecht, M. (2007) Functional evaluation of domain-domain interactions and human protein interaction networks. *Bioinformatics* 23:859-865.

Schueler-Furman, Ora; Glick, Eitan; Segovia, José; Linial, Michal (2006) Is GAS1 a co-receptor for the GDNF family of ligands? *Trends Pharmacol.Sci.*, 27, 2, 72-77.

Sekowska, A, Déneraud, V, Ashida, H, Michoud, K, Haas, D, Yokota, A, Danchin, A. (2004) Bacterial variations on the methionine salvage pathway. *BMC Microbiol* 4: 9.

Serafini-Fracassini D, Della Mea M, Tasco G, Casadio R, Del Duca S. (2009) Plant and animal transglutaminases: do similar functions imply similar structures? *Amino Acids* 36, 643-657.

Servitja, JM; Pignatelli, M; Maestro, MA; Cardalda, C; Boj, SF; Lozano, J; Blanco, E; Lafuente, A; McCarthy, MI; Sumoy, L; Guigó, R; Ferrer, J. (2009) Hnf1alpha (MODY3) controls tissue-specific transcriptional

programs and exerts opposed effects on cell growth in pancreatic islets and liver. *Mol Cell Biol.* Jun; 29(11):2945-59.

Sillitoe, M. Dibley, J. Bray, S. Addou & Christine Orengo (2005) Assessing Strategies For Improved Superfamily Recognition in Genome Annotation I. *Protein Science.* 14:1800-10.

Sing T, Low AJ, Beerenwinkel N, Sander O, Cheung PK, Domingues FS, Büch J, Däumer M, Kaiser R, Lengauer T, Harrigan PR. (2007) Predicting HIV coreceptor usage on the basis of genetic and clinical covariates. *Antivir Ther* 12:1097-1106.

Sing, T.; Sander, O; Beerenwinkel, N; Lengauer, T. (2005) ROCr: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940-1.

Sing, T; Svicher, V; Beerenwinkel, N; Ceccherini-Silberstein, F; Däumer, M; Kaiser, R; Walter, H; Korn, K; Hoffmann, D; Oette, M; Rockstroh, JK; Fätkenheuer, G; Perno, C-F and Lengauer, T. (2005) Characterization of Novel HIV Drug Resistance Mutations Using Clustering, Multidimensional Scaling and SVM-Based Feature Ranking, in: *Knowledge Discovery in Databases: PKDD 2005* M.J. Alipio, et al., Editors, Lecture Notes in Computer Science No. 3721, 285-296 Springer Verlag.

Smeitink, JAM, Elpeleg, O, Antonicka, H, Diepstra, H, Saada, A, Smits, P, Sasarman, F, Vriend, G, Jacob-Hirsch, J, Shaag, A, Rechavi, G, Welling, B, Horst, J, Rodenburg, RJ, van den Heuvel, B, Shoubbridge, EA. (2006) Distinct Clinical Phenotypes Associated with a Mutation in the Mitochondrial Translation Elongation Factor EFTs *The American Journal of Human Genetics*, 79; 869–877.

Smidtas S, Yartseva A, Schachter V, Kepes F. (2006) Model of interactions in biology and application to heterogeneous network in yeast. *C R Biol*;3 29(12):945-52. Epub 2006 Aug 7

Smidtas, S, Schachter, V, Képès, F. (2006) The adaptive filter of the yeast galactose pathway *J Theor Biol*; 242(2):372-81.

Sommer I, Müller O, Domingues FS, Sander O, Weickert J, Lengauer T. (2007) Moment invariants as shape recognition technique for comparing protein binding sites. *Bioinformatics* 23:3139-3146.

Soro S, Orecchia A, Morbidelli L, Lacal PM, Morea V, Ballmer-Hofer K, Ruffini F, Ziche, M, D'Atri S, Zambruno G, Tramontano A, Failla CM. (2008) A proangiogenic peptide derived from vascular endothelial growth factor receptor-1 acts through alpha5beta1 integrin *Blood.*, 111(7):3479-88

Soro, S., and Tramontano, A. (2005) Function Prediction at CASP6 *Proteins*, S7, 214-224.

Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., Crosby, M. A., Rasmussen, M. D., Roy, S., Deoras, A. N., Ruby, J. G., Brennecke, J., Curators, H. F., Project, B. D., Hodges, E., Hinrichs,

A. S., Caspi, A., Paten, B., Park, S. W., Han, M. V., Maeder, M. L., Polansky, B. J., Robson, B. E., Aerts, S., van Helden, J., Hassan, B., Gilbert, D. G., Eastman, D. A., Rice, M., Weir, M., Hahn, M. W., Park, Y., Dewey, C. N., Pachter, L., Kent, W. J., Haussler, D., Lai, E. C., Bartel, D. P., Hannon, G. J., Kaufman, T. C., Eisen, M. B., Clark, A. G., Smith, D., Celniker, S. E., Gelbart, W. M., Kellis, M., Matthews, B. B., Schroeder, A. J., Sian Gramates, L., St Pierre, S. E., Roark, M., Wiley Jr, K. L., Kulathinal, R. J., Zhang, P., Myrick, K. V., Antone, J. V., Yu, C., Park, S. & Wan, K. H. (2007). Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* 450, 219-232. Pubmed 17994088.

Stranger, B., Forrest, M., Clark, A., Minichiello, M., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S., Tavare, S., Deloukas, P., Dermitzakis, E. (2005) Genome-Wide Associations of Gene Expression Variation in Humans *PLoS Genet.*, 6:e78.

Sultan, Marc; Schulz, Marcel H.; Richard, Hugues; Magen, Alon; Klingenhoff, Andreas; Scherf, Matthias; Seifert, Martin; Borodina, Tatjana; Soldatov, Aleksey; Parkhomchuk, Dmitri; Schmidt, Dominic; O'Keefe, Sean; Haas, Stefan; Vingron, Martin; Lehrach Hans and Yaspo, Marie-Laure (2008) A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science, New Series*, Vol. 321, No. 5891, 956-960.

Sutton L, Kostareli E, Darzentas N, Hadzidimitriou A, et al. (2009) Extensive intraclonal diversification in a subgroup of chronic lymphocytic leukemia patients with stereotyped IGHV4-34 receptors: implications for ongoing interactions with antigen. [submitted]

Svicher, Valentina; Sing, Tobias; Santoro, Maria Mercedes; Forbici, Federica; Rodríguez-Barrios, Fátima; Bertoli, Ada; Beerenwinkel, Niko; Bellocchi, Maria Concetta; Gago, Federico; d'Arminio Monforte, Antonella; Antinori, Andrea; Lengauer, Thomas; Ceccherini-Silberstein, Francesca; and Perno, Carlo Federico (2006) Involvement of Novel Human Immunodeficiency Virus Type 1 Reverse Transcriptase Mutations in the Regulation of Resistance to Nucleoside Inhibitors. *J. Virol.* July 80: 7186-7198.

Takeuchi, Akiko; Schmitt, David; Chapple, Charles; Babaylova, Elena; Karpova, Galina; Guigo, Roderic; Krol, Alain; Allmang, Christine (2009) A short motif in *Drosophila* SECIS Binding Protein 2 provides differential binding affinity to SECIS RNA hairpins. *Nucleic Acids Res.*, 37, 7, 2126-2141.

Talavera, David; Laskowski, Roman A. and Thornton, Janet M. (2009) WSSas: a web service for the annotation of functional residues through structural homologues. *Bioinformatics* 25: 1192-1194; doi:10.1093/bioinformatics/btp116

Talavera, D; Taylor, MS and Thornton, JM (2009) The (non)-malignancy of cancerous amino acidic substitutions. *PROTEINS: Structure, Function, and Bioinformatics*. In press

- Tameling, Vladimir I.L.; Vossen, Jack H.; Albrecht, Mario; Lengauer, Thomas; Berden, Jan A.; Haring, Michel A.; Cornelissen, Ben J.C. and Takken, Frank L.W. (2006) Mutations in the NB-ARC Domain of I-2 That Impair ATP Hydrolysis Cause Autoactivation. *Plant Physiol.*140: 1233-1245. doi:10.1104/pp.105.073510
- Tassoni A, Franceschetti M, Tasco G, Casadio R, Bagni N. (2007) Cloning, functional identification and structural modelling of *Vitis vinifera* S-adenosylmethionine decarboxylase. *J Plant Physiol*; 164(9):1208-19.
- Taylor M, Valdar W, Kumar A, Flint J, Mott R (2007) Management, presentation and interpretation of genome scans using GSCANDB. *Bioinformatics* 23, 1545 – 1549.
- Theodosiou T, Darzentas N, Angelis L, Ouzounis CA. 2008) PuReD-MCL: a graph-based PubMed document clustering methodology. *Bioinformatics*. 24(17):1935-1941.
- Thomas-Chollier, Morgane; Sand, Olivier; Turatsinze, Jean-Valéry; Janky, Rekin's; Defrance, Matthieu; Vervisch, Eric; Brohée, Sylvain and van Helden, Jacques (2008) RSAT: regulatory sequence analysis tools. *Nucl. Acids Res.* 36: W119-W127; doi:10.1093/nar/gkn304
- Tiffin N, Adie E, Turner F, Brunner HG, van Driel MA, Oti M, López-Bigas N, Ouzounis CA, Pérez-Irratxeta C, Andrade-Navarro MA, Adeyemo A, Patti ME, Semple CA, Hide W. (2006) Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucl. Acids Res.* 34, 3067-3081.
- Torchala, M., Hoffmann, M. (2007) IA, database of known ligands of aminoacyl-tRNA synthetases *J Comput Aided Mol Des* 21:523—525.
- Tordai H , Nagy A , Farkas K , Bányai L and Patthy L. (2005) Modules, multidomain proteins and organismic complexity. *FEBS J* 272, 5064-78.
- Tordai H, Nagy A, Farkas K, Bányai L, Hegyi H, Patthy L. (2006) MisPred – Database for mispredicted and abnormal proteins. <http://mispred.enzim.hu/index.html>.
- Torrance JW, Macarthur MW, Thornton JM. (2008) Evolution of binding sites for zinc and calcium ions playing structural roles. *Proteins*. 71(2):813-30.
- Tramontano, A. (2007) Worth the effort: an account of the Seventh meeting of the worldwide critical assessment of Techniques for Protein Structure Prediction. *FEBS J.*, 274, 1651-1654.
- Tress, M; Bodenmiller, B; Aebersold, R; Valencia, A (2008) Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biology* 9 (11) R162.
- Tress M, Cheng J, Baldi P, Joo K, Lee J, Seo JH, Lee J, Baker D, Chivian D, Kim D, Ezkurdia I. (2007) Assessment of predictions

submitted for the CASP7 domain prediction category. *Proteins*. 69 Suppl 8:137-51.

Tress, M., Cozzetto, D., Tramontano, A. and Valencia, A. (2006) An analysis of the Sargasso Sea Resource and the consequences of database composition *BMC Bioinformatics* 7, 213.

Tress, M., de Juan, D., Graña, O., Gómez, M., Gómez-Puertas, P., González, J., López, G. and Valencia, A. (2005) Scoring docking models with evolutionary information *Proteins: Structure, Function, and Bioinformatics*, 60, 275-280.

Tress, M., Ezkurdia, I., Graña, O., López, G., and Valencia, A. (2005) Assessment of predictions submitted for the CASP6 comparative modeling category *Proteins*, 61(S7), 27-45.

Tress, M.L., Martelli, P.L., Frankish, A., Reeves, G.A., Wesselink, J.J., Yeats, C., Ólason, P.Í., Albrecht, M., Hegyi, H., Giorgetti, A., Raimondo, D., Lagarde, J., Laskowski, R.A., López, G., Sadowski, M.I., Watson, J.D., Fariselli, P., Rossi, I., Nagy, A., Kai, W., Størling, Z., Orsini, M., Assenov, Y., Blankenburg, H., Huthmacher, C., Ramírez, F., Schlicker, A., Denoeud, F., Jones, P., Kerrien, S., Orchard, S., Antonarakis, S.E., Reymond, A., Birney, E., Brunak, S., Casadio, R., Guigó, R., Harrow, J., Hermjakob, H., Jones, D.T., Lengauer, T., Orengo, C.A., Pathy, L., Thornton, J.M., Tramontano, A., Valencia, A. (2007) The implications of alternative splicing in the ENCODE protein complement. *Proc.Natl.Acad.Sci.* 104:5495-5500.

Tress, M.L., Wesselink, J.J., Frankish, A. et al. (2008), Determination and validation of principal gene products, *Bioinformatics* 24 (1), 11.

Turatsinze, J.V., Thomas-Chollier, M., Defrance, M. and van Helden, J. (2008) Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat Protoc*, 3, 1578-1588. Pubmed 18802439.

Tuupanen, S, Turunen, M, Lehtonen, R, Hallikas, O, Vanharanta, S, Kivioja, T, Björklund, M, Wei, G, Yan, J, Niittymäki, I, Mecklin, J-P, Järvinen, H, Ristimäki, A, Di-Bernardo, M, East, P, Carvajal-Carmona, L, Houlston, RS, Tomlinson, I, Palin, K, Ukkonen, E, Karhu, A, Taipale, J and Aaltonen, LA (2009) The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nature Genetics*, in press.

Valdar, W; Holmes, CC; Mott, R; Flint, J. (2009) Mapping in Structured Populations by Resample Model Averaging. *Genetics*, May 2009; doi:10.1534/genetics.109.100727

Valdar W, Solberg LC, Gauguier D, Cookson WO, Rawlins NJ, Mott R, Flint J. (2006) Genetic and environmental effects on complex traits in mice. *Genetics* 174(2):959-84.

Valdar W, Solberg Leah, Gauguier D, Burnett S, Klenerman P, Cookson WOC, Taylor MS, Rawlins N, Mott R, Flint J. (2006) Genome-wide

genetic association of complex traits in heterogenous stock mice *Nature Genetics* 38(8):879-87.

Valdar, W., Solberg, L., Gauguier, D., Cookson, W., Rawlins, N., Mott, R., Flint, J. (2006) Genetic and environmental effects on complex traits in mice. *Genetics*. 174(2):959-84. Epub 2006 Aug 3.

Valencia, A. and the BioSapiens Consortium. (2005) BioSapiens: a European network for integrated genome annotation. *Eur J Hum Genet* 13: 994-997.

Valencia, Alfonso (2005) Automatic annotation of protein function. *Curr.Opin.Struct.Biol.*, 15, 3, 267-274.

Valentonyte, Ruta; Hampe, Jochen; Huse, Klaus; Rosenstiel, Philip; Albrecht, Mario; Stenzel, Annette; Nagy, Marion; Gaede, Karoline I; Franke, Andre; Haesler, Robert; Koch, Andreas; Lengauer, Thomas; Seegert, Dirk; Reiling, Norbert; Ehlers, Stefan; Schwinger, Eberhard; Platzer, Matthias; Krawczak, Michael; Müller-Quernheim, Joachim; Schürmann, Manfred and Schreiber, Stefan (2005) Sarcoidosis is associated with a truncating splice site mutation in *BTNL2*. *Nat.Genet.*, 37, 4, 357-364.

van Ooijen, Gerben; Mayr, Gabriele; Albrecht, Mario; Cornelissen, Ben J. C. and Takken, Frank L.W. (2008) Transcomplementation, but not Physical Association of the CC-NB-ARC and LRR Domains of Tomato R Protein Mi-1.2 is Altered by Mutations in the ARC2 Subdomain. *Mol Plant* 1: 401-410; doi:10.1093/mp/ssn009

van Ooijen, Gerben; Mayr, Gabriele; Kasiem, Mobien M. A.; Albrecht, Mario; Cornelissen, Ben J. C. and Takken, Frank L.W. (2008) Structure–function analysis of the NB-ARC domain of plant disease resistance proteins. *J. Exp. Bot.* 59: 1383-1397; doi:10.1093/jxb/ern045

Varshavsky, Roy; Linial, Michal; Horn, David (2005) COMPACT: A Comparative Package for Clustering Assessment Parallel and Distributed Processing and Applications. *ISPA 2005 Workshops*, 159-167.

Vassura M, Margara L, Di Lena P, Medri F, Fariselli P, Casadio R (2007) Fault tolerance for large scale protein 3D reconstruction from contact maps- *Lect Notes Comp Sci LNCS 4645*: 25-37.

Vassura M, Margara L, Di Lena P, Medri F, Fariselli P, Casadio R. (2008) FT-COMAR: fault tolerant three-dimensional structure reconstruction from protein contact maps. *Bioinformatics* 24:1313-1315.

Vassura M, Margara L, Di Lena P, Medri F, Fariselli P, Casadio R. (2008) Reconstruction of 3D structures from protein contact maps. *IEEE/ACM Trans Comput Biol Bioinform* 5:357-367.

Vassura M, Margara L, Fariselli P, Casadio R. (2007) A graph theoretic approach to protein structure selection- *Lect Notes Comp Sci* 4578: 497-504.

Vassura M, Margara L, Fariselli P, Casadio R. (2009) A graph theoretic approach to protein structure selection. *Artif Intell Med* 45:229-237.

Vassura M, Margara L, Medri F, Di Lena P, Fariselli P, Casadio R. (2007) Reconstruction of 3D structures from protein contact maps- *Lect Notes Comp Sci LNCS* 4463: 578-589.

Vassura, Marco; Margara, Luciano; Di Lena, Pietro; Medri, Filippo; Fariselli, Piero; Casadio, Rita (2007) Fault Tolerance for Large Scale Protein 3D Reconstruction from Contact Maps. *Algorithms in Bioinformatics, 2007*, 25-37.

Vingron M, Brazma A, Coulson R, van Helden J, Manke T, Palin K, Sand O., Ukkonen E. (2009). Integrating sequence, evolution and functional genomics in regulatory genomics. *Genome Biol.* 10(1): 202.

von Grotthuss, Marcin; Plewczynski, Dariusz; Vriend, Gerd; Rychlewski, Leszek. (2008) 3D-Fun: predicting enzyme function from structure. *Nucleic Acids Res.* 36(Web Server issue):W303-7.

von Grotthuss, Marcin; Plewczynski, Dariusz; Vriend, Gert and Rychlewski, Leszek (2008) 3D-Fun: predicting enzyme function from structure. *Nucl. Acids Res.* 36: W303-W307; doi:10.1093/nar/gkn308

von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Krüger B, Snel B and Bork P. (2007) STRING 7—recent developments in the integration and prediction of protein interactions, *Nucleic Acids Research*, 35, D358-D362.

von Mering, C; Hugenholtz, P; Raes, J; Tringe, S.G.; Doerks, T.; Jensen, L.J; Ward, N and Bork, P (2007) Quantitative Phylogenetic Assessment of Microbial Communities in Diverse Environments. *Science* 23 February 2007: 1126-1130. DOI: 10.1126/science.1133420

Walter, MC; Rattei, T; Arnold, R; Güldener, U; Münsterkötter, M; Nenova, K; Kastenmüller, G; Tischler, P; Wölling, A; Volz, A; Pongratz, N; Jost, R; Mewes, HW; Frishman, D. (2009) PEDANT covers all complete RefSeq genomes. *Nucleic Acids Res.* 37 (Database issue), D408-D411.

Watson JD, Laskowski RA & Thornton JM (2005) Predicting Protein Function From Sequence And 3D Structural Data. *Curr. Opin. Struct. Biol.* 15 (3), 275 – 284.

Watson JD, Sanderson S, Ezersky A, Savchenko A, Edwards A, Orengo C, Joachimiak A, Laskowski RA, Thornton JM (2007) Towards Fully Automated Structure Based Function Prediction in Structural Genomics: a Case Study. *J Mol Biol.* 367(5):1511-22. Epub 2007 Jan 30.

Weinhold N, Sander O, Domingues FS, Lengauer T, Sommer I. Local Function Conservation in Sequence and Structure Space. *PLoS Comput Biol.* 2008 4:e1000105.

Welsch, Christoph; Albrecht, Mario; Maydt, Jochen; Herrmann, Eva; Welker, Martin Walter; Sarrazin, Christoph; Scheidig, Axel; Lengauer, Thomas; Zeuzem Stefan (2007) Structural and functional comparison of the non-structural protein 4B in flaviviridae. *J.Mol.Graph.Model.*, 26, 2, 546-557.

Welsch, C; Domingues, FS; Susser, S; Antes, I; Hartmann, C; Mayr, G; Schlicker, A; Sarrazin, C; Albrecht, M; Zeuzem, S; Lengauer, T.(2008) Molecular basis of telaprevir resistance due to V36 and T54 mutations in the NS3-4A protease of the hepatitis C virus. *Genome Biol.* 9:R16.

Wong P., Althammer S., Hildebrand A., Kirschner A., Pagel P., Geissler B., Smialowski P., Bloechl, M., Oesterheld M., Schmidt T., Strack N., Theis, F., Ruepp A., Frishman, D. (2008) An evolutionary and structural characterization of mammalian protein complex organization. *BMC Genomics*, 9(1), 629.

Wyrwicz LS, Rychlewski L. (2007) Herpes glycoprotein gL is distantly related to chemokine receptor ligands. *Antiviral Research* 75 (1), 83-86.

Wyrwicz, L., Gaj, P., Hoffmann, M., Rychlewski, L., Ostrowski, J. (2007) A common cis-element in promoters of protein synthesis and cell cycle genes. *Acta Biochim Pol.* 54 (1), 89–98.

Wyrwicz, L., Koczyk, G., Rychlewski, L., Plewczyński, D. (2007) ProteinSplit: splitting of multi-domain proteins using prediction of ordered and disordered regions in protein sequences for virtual structural genomics. *Journal of Physics: Condensed Matter* 28; 285222.

Wyrwicz, L., Rychlewski, L. (2007) Fold recognition insights into function of Herpes ICP4 protein. *Acta Biochim Pol.* 54 (3), 551–559.

Wyrwicz, L., Rychlewski, L. (2007) Identification of Herpes TATT-binding protein. *Antiviral Res.* 75, 167 – 172.

Yap, YeeLeng; Zhang, XueWu; Andonov, Anton; He, RunTao (2005) Structural analysis of inhibition mechanisms of Aurintricarboxylic Acid on SARS-CoV polymerase and other proteins *Computational Biology and Chemistry*, 29, 3, 212-219.

Yeats C, Lees J, Reid A, Kellam P, Martin N, Liu X, Orengo C (2008) Gene3D: Comprehensive structural and functional annotation of genomes. *Nucleic Acid Res* 36: D414-8

Yeats, C., Maibaum, M., Marsden, R., Dibley, M., Lee, D., Addou, S., and Orengo, C. (2005) Gene3D: modelling protein structure, function and evolution. *Nucleic Acids Res.*, 34:D281 - D284.

Zachariae, Ulrich; Schneider, Robert; Velisetty, Phanindra; Lange, Adam; Seeliger, Daniel; Wacker, Sören J.; Karimi-Nejad, Yasmin; Vriend, Gert; Becker, Stefan; Pongs, Olaf; Baldus, Marc; de Groot, Bert L.(2008) The Molecular Mechanism of Toxin-Induced Conformational Changes in a Potassium Channel: Relation to C-Type Inactivation *Structure* 16, 5, 747-754.

Zemojtel, T., Penzkofer, T., Duchniewicz, M. and Zwartkruis, F.J.T. (2006) HRap1B-retro: a novel human processed rap1B gene blurs the picture? *Leukemia*, 20, 1,145-6.

Zemojtel, Tomasz; Kolanczyk, Mateusz; Kossler, Nadine; Stricker, Sigmar; Lurz, Rudi; Mikula, Ivan; Duchniewicz, Marlena; Schuelke, Markus; Ghafourifar, Pedram; Martasek, Pavel; Vingron, Martin; Mundlos, Stefan (2006) Mammalian mitochondrial nitric oxide synthase: Characterization of a novel candidate *FEBS Lett.*, 580, 2, 455-462.

Zemojtel, Tomasz; Fröhlich, Andreas; Palmieri, M. Cristina; Kolanczyk, Mateusz; Mikula, Ivan; Wyrwicz, Lucjan S.; Wanker, Erich E.; Mundlos, Stefan; Vingron, Martin; Martasek, Pavel; Durner, Jörg (2006) Plant nitric oxide synthase: a never-ending story? *Trends Plant Sci.*, 11, 11, 524-525.

Zemojtel, Tomasz; Kielbasa, Szymon M.; Arndt, Peter F.; Chung, Ho-Ryun; Vingron, Martin (2009) Methylation and deamination of CpGs generate p53-binding sites on a genomic scale. *Trends in Genetics*, 25, 2, 63-66.

Zhu H, Domingues FS, Sommer I, Lengauer T. (2006) NOXclass: prediction of protein-protein interaction types *BMC Bioinformatics*, 7:27.

Zhu H, Sommer I, Lengauer T, Domingues FS. Alignment of Non-Covalent Interactions at Protein-Protein Interfaces. *PLoS ONE*. 2008 3:e1926