Argomenti del corso:

Introduzione alla bioinformatica

Strumenti della bioinformatica

Ricerca su banche dati

Allineamenti di sequenze

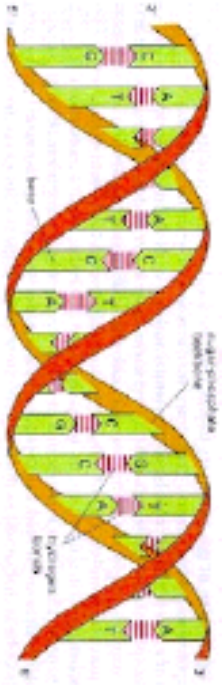
Analisi struttura proteine

Famiglie proteine

Alberi filogenetici

Metodi per la predizione della struttura

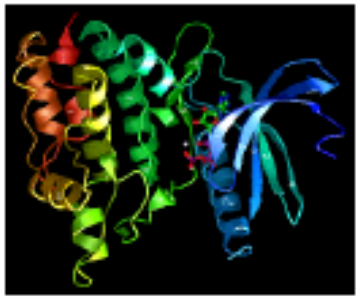
Central Dogma



Transcription

mRNA

Translation

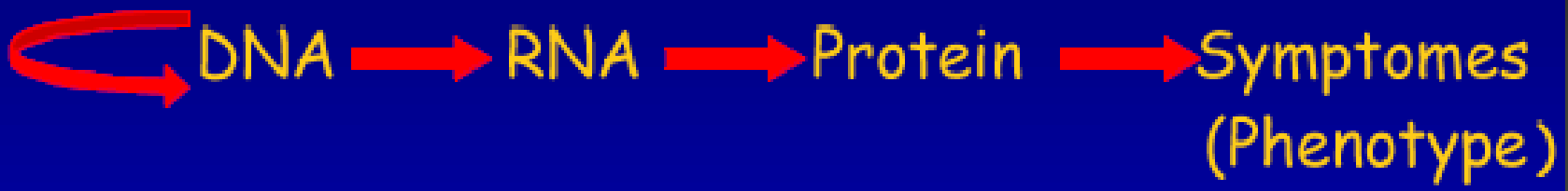


Protein

Gene (DNA)

Cells express **different** subset of the genes in different tissues and under different conditions

Central Paradigm of Molecular Biology



What is Bioinformatics ?

Bioinformatics is the use of computers for the acquisition, management, and analysis of biological information.

It incorporates elements of molecular biology, computational biology, database computing, and the Internet…

… bioinformatics is clearly a multi-disciplinary field including: computer systems management networking, database design, computer programming, molecular biology

From Using Computers for Molecular Biology, Stuart M. Brown, PhD, RCR, NYU Medical Center

Bioinformatics is a multifaceted discipline combining many scientific fields including computational biology, statistics, mathematics, molecular biology and genetics (Fenstermacher, 2005, p. 440).

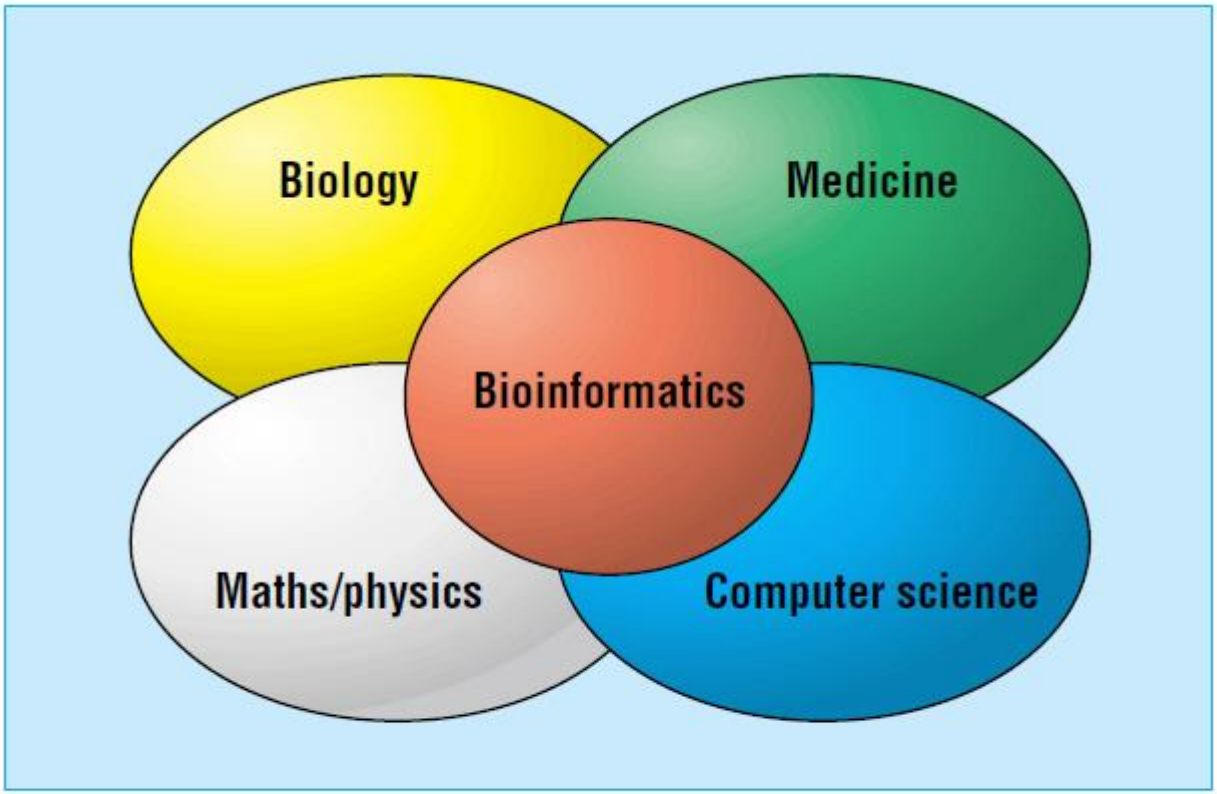


Fig 1 Interaction of disciplines that have contributed to the formation of bioinformatics

Bioinformatics: Origins & Definitions

Bioinformatics - a definition[1]

(Molecular) **bio** – informatics: bioinformatics is conceptualising biology in terms of molecules (in the sense of physical chemistry) and applying "***informatics techniques***" (derived from disciplines such as applied maths, computer science and statistics) to ***understand*** and ***organise*** the ***information*** associated with these molecules, on a ***large scale***. In short, bioinformatics is a management information system for molecular biology and has many ***practical applications***.

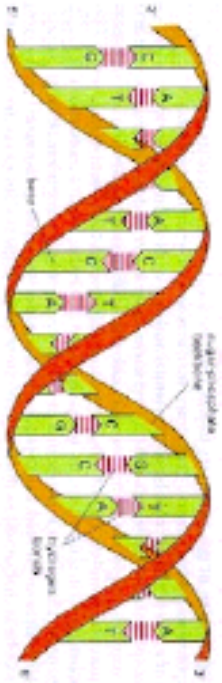
[1] As submitted to the Oxford English Dictionary

Bioinformatics has many definitions

… *the study of how information is represented and analyzed in biological systems, starting at the molecular level … concerned with understanding how basic biological systems conspire to create molecules, organelles, living cells, organs, and entire organisms* (Altman & Mooney, 2006, p. 763)

… *application of tools of computation and analysis to the capture and interpretation of biological data* (Bayat, 2003, p. 1018)

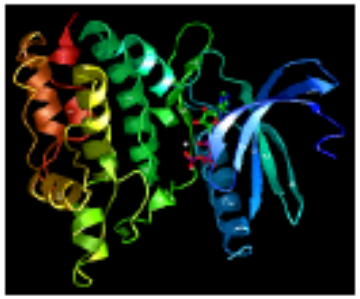
Central Dogma



Transcription

mRNA

Translation

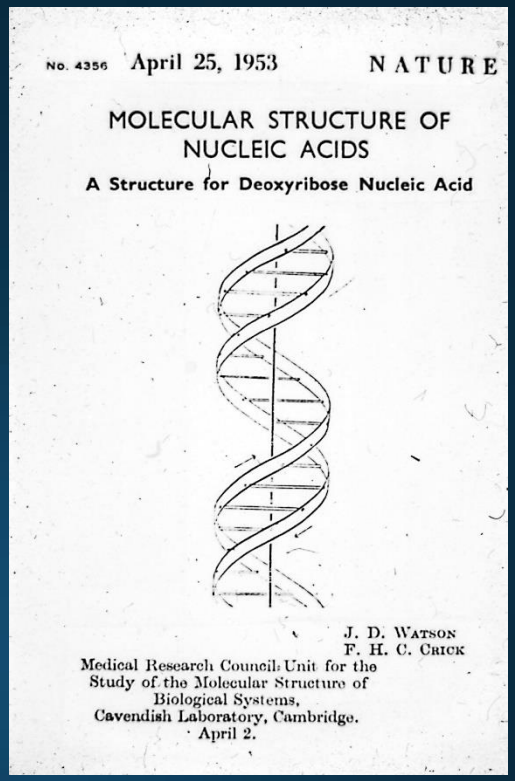


Protein

Gene (DNA)

Cells express **different** subset of the genes in different tissues and under different conditions

DNA is the nature's universal information storage medium



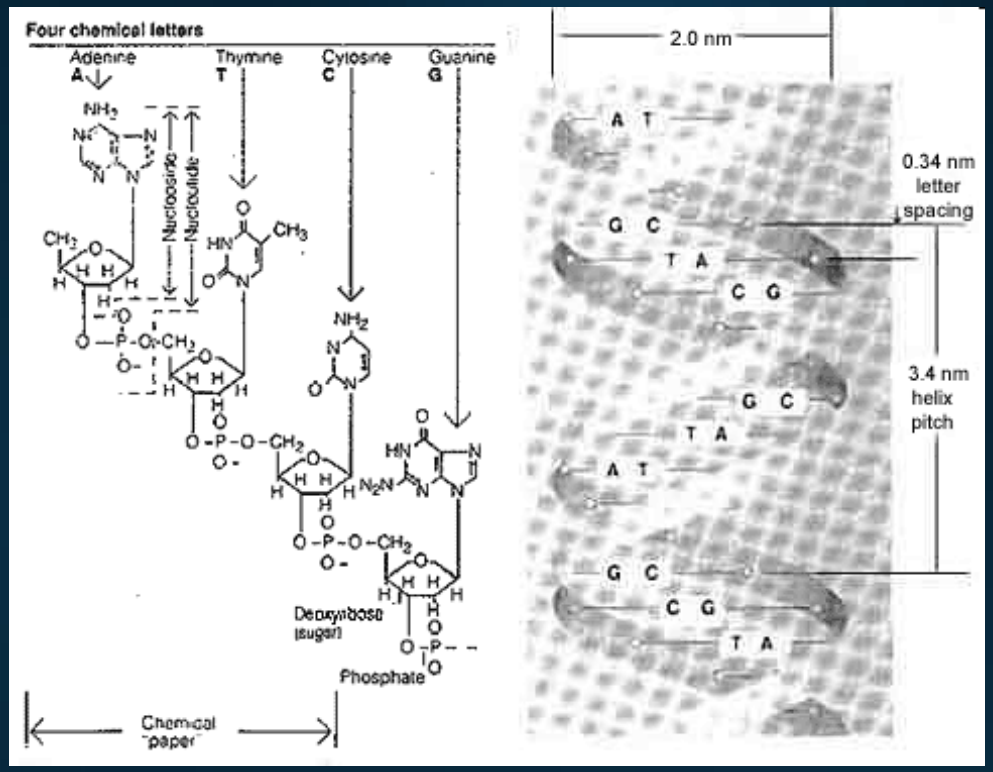
No. 4356 April 25, 1953 NATURE

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

J. D. Watson
F. H. C. Crick

Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems, Cavendish Laboratory, Cambridge. April 2.



… increasingly, biological research relies on information science

The Human Genome Project

- Produced the human genome sequence
- Spawned a new field: genomics
- Spurred new technologies
- And now provides us an unparalleled opportunity to apply new knowledge, technologies, and approaches to health care

Guttmacher (2009)

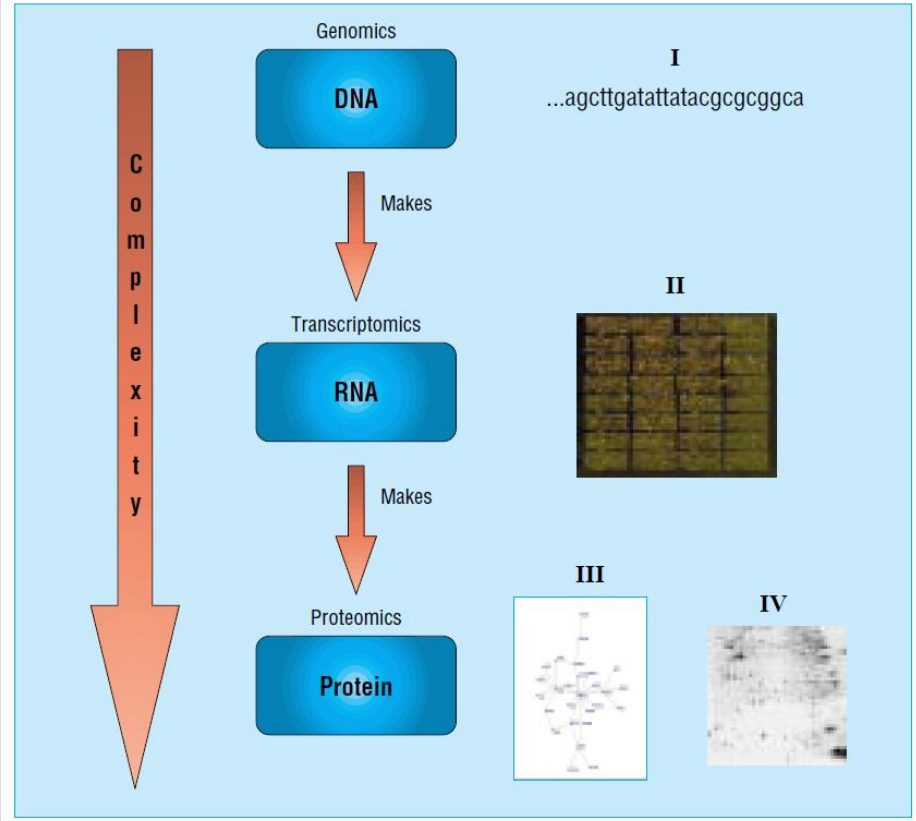


Fig 4 Schematic diagram representing complexity of genomic data processing. Analysis and interpretation of biological data considers information at every level from the genome (total genetic content) to the proteome (total protein content) and transcriptome (total messenger RNA content) of the cell. The images numbered I-IV to the right of the diagram represent relevant examples of DNA (image I is base pair nucleotides); RNA (image II is a microarray showing levels of gene expression); and protein (image III is a structure of a single protein; image IV is a two dimensional gel electrophoresis showing separation of all proteins of a cell—each spot corresponds to a different protein chain)

Bioinformatics supports "-omics" research

…from Bayat (2002), p 1020.

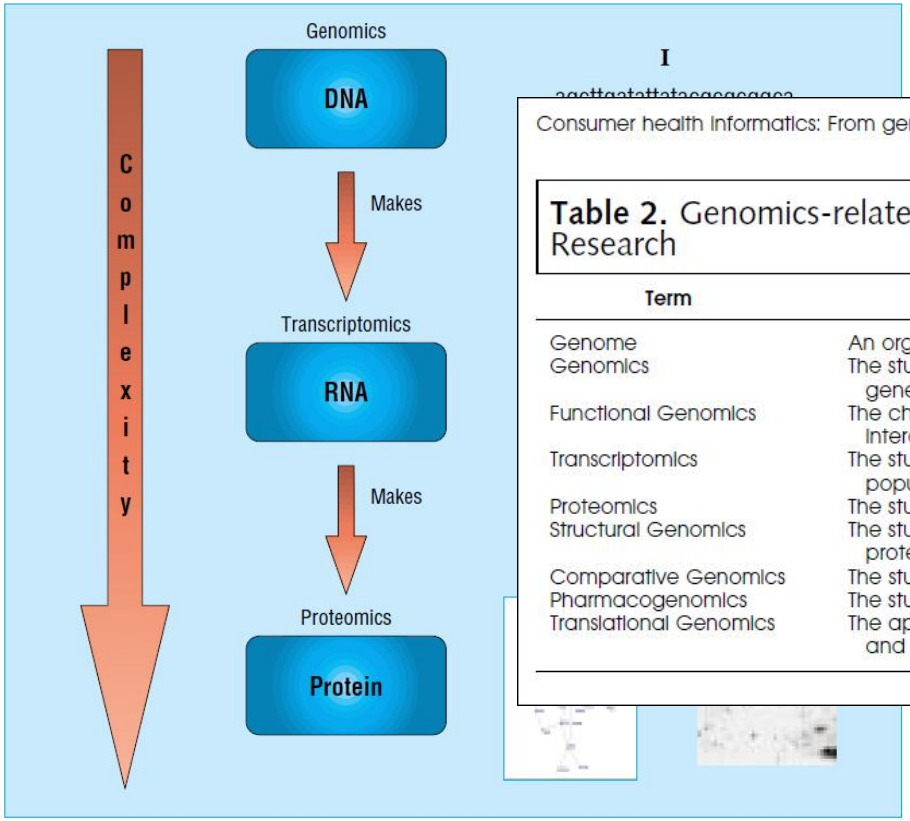


Fig 4 Schematic diagram representing complexity of genomic data processing. Analysis and interpretation of biological data considers information at every level from the genome (total genetic content) to the proteome (total protein content) and transcriptome (total messenger RNA content) of the cell. The images numbered I-IV to the right of the diagram represent relevant examples of DNA (image I is base pair nucleotides); RNA (image II is a microarray showing levels of gene expression); and protein (image III is a structure of a single protein; image IV is a two dimensional gel electrophoresis showing separation of all proteins of a cell—each spot corresponds to a different protein chain)

…from Bayat (2002), p 1020.

Table 2. Genomics-related Terminology and Description of Related Research

Term	Description of Related Research
Genome	An organism's total set of genes.
Genomics	The study of an organism's entire genome, including the interactions between genes and the interactions between genes and environment.
Functional Genomics	The characterization of genes to determine their structure, function, and interactions.
Transcriptomics	The study of the expression level of mRNA (transcript) in a given cell population under given conditions.
Proteomics	The study of proteins to establish their structure and function.
Structural Genomics	The study and subsequent development of 3D structures of one or more proteins from each protein family.
Comparative Genomics	The study of the relationship or comparison of genomes across species.
Pharmacogenomics	The study of the influence of genetic variation on drug response.
Translational Genomics	The application or translation of genomics and genomics-related discovery and technology into clinically useful information and tools.

…from McDaniel, Schutte, & Keller (2008), p. 220

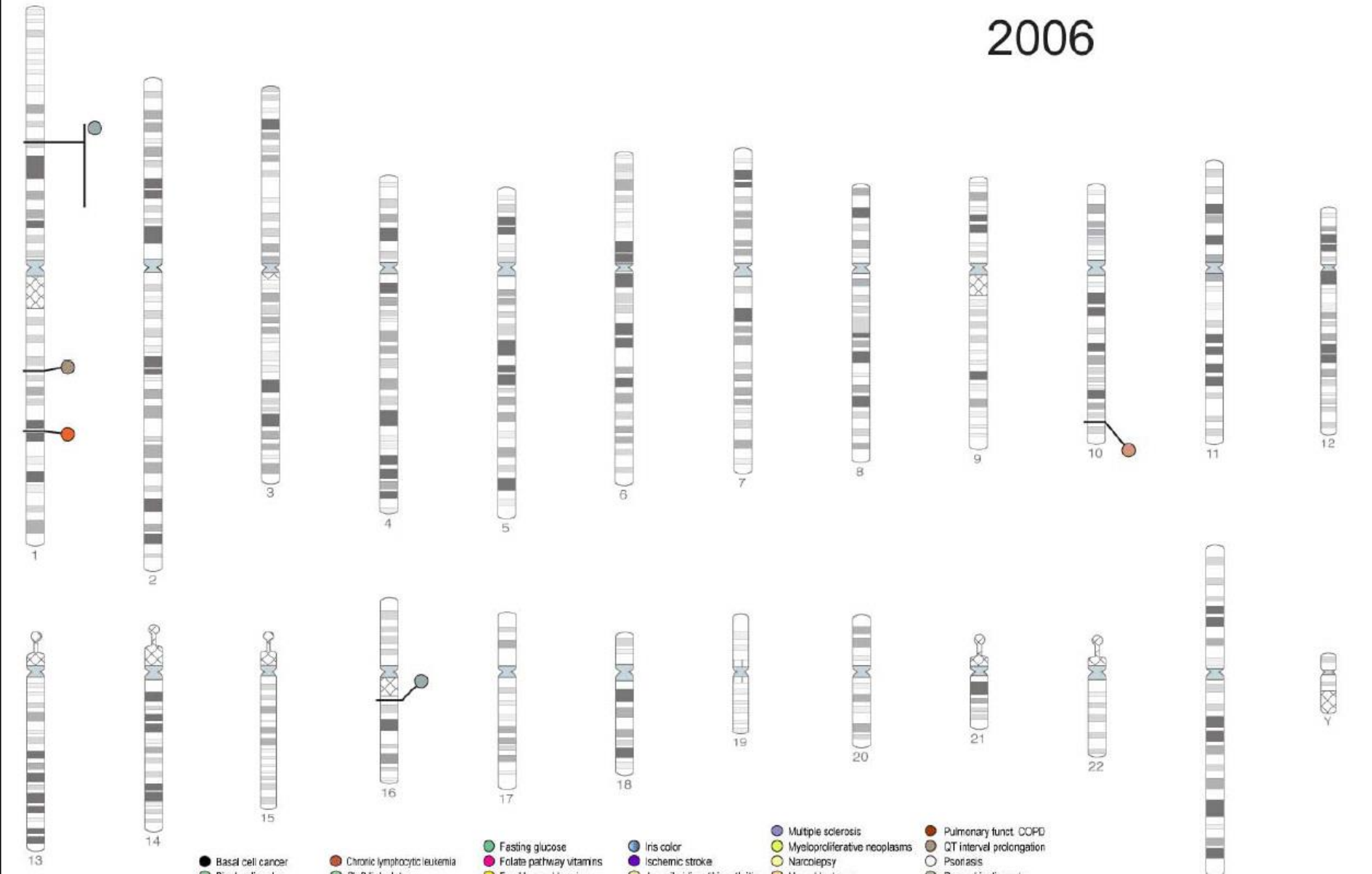
Central Paradigm of Molecular Biology

Bioinformatics Data

- Bioinformatics deals with any type of data that is of interest to biologists
  - DNA and protein sequences
  - Gene expression (microarray)
  - Raw data collected from field or laboratory experiment
  - Images, virtual models, Software
  - Articles from literature and databases of citations
- Each type of data can exist in many incompatible computer formats
- The analysis of DNA sequence data has come to dominate the field of bioinformatics, but the term can be applied to any type of biological data that can be recorded as numbers or images and handled by computers

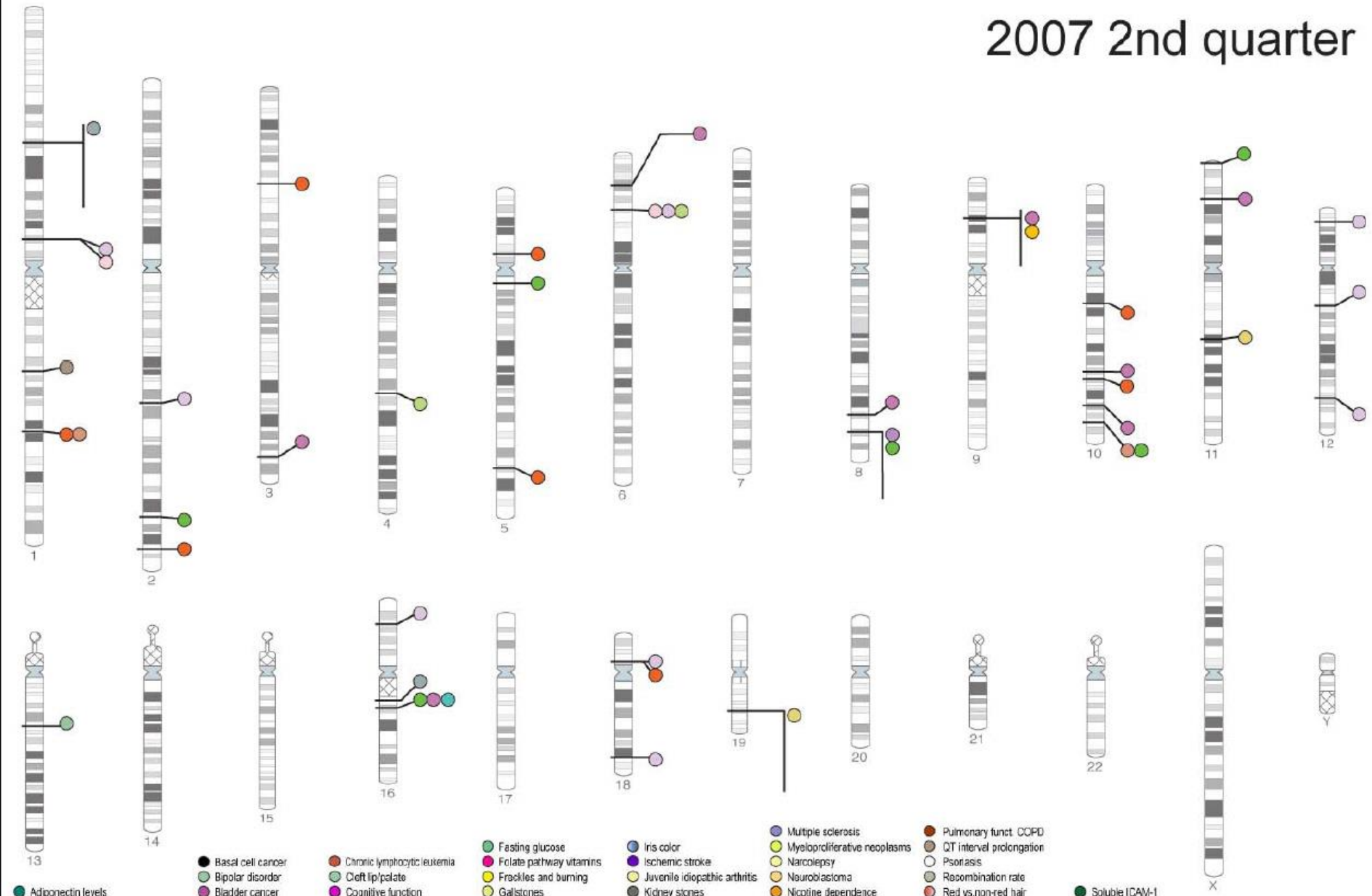
From Using Computers for Molecular Biology, Stuart M. Brown, PhD, RCR, NYU Medical Center

2006

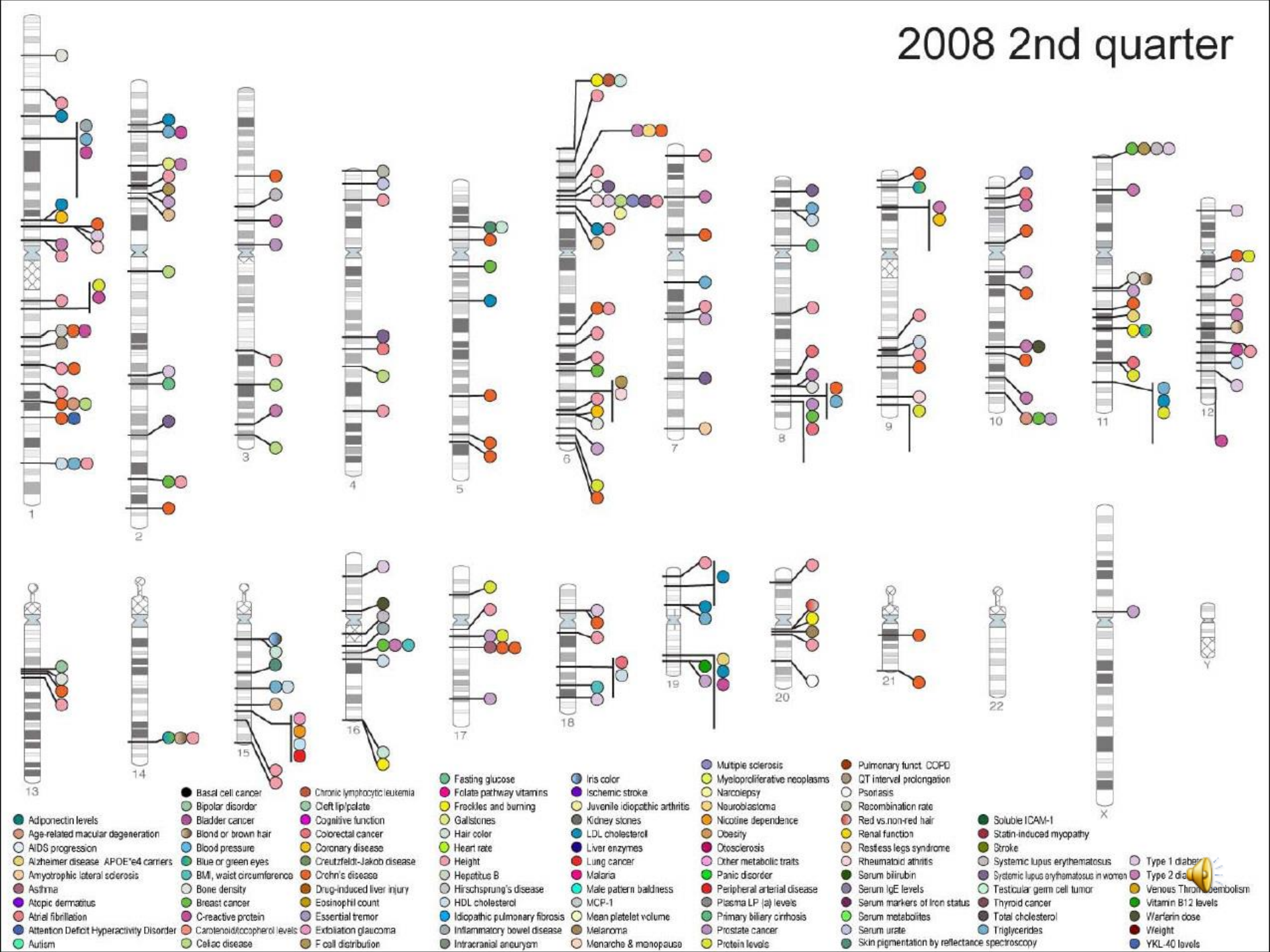


- Adiponectin levels
- Age-related macular degeneration
- AIDS progression
- Alzheimer disease APOE*e4 carriers
- Amyotrophic lateral sclerosis
- Asthma
- Atopic dermatitus
- Atrial fibrillation
- Attention Deficit Hyperactivity Disorder
- Autism
- Basal cell cancer
- Bipolar disorder
- Bladder cancer
- Blond or brown hair
- Blood pressure
- Blue or green eyes
- BMI, waist circumference
- Bone density
- Breast cancer
- C-reactive protein
- Carotenoid/tocopherol levels
- Celiac disease
- Chronic lymphocytic leukemia
- Cleft lip/palate
- Cognitive function
- Colorectal cancer
- Coronary disease
- Creutzfeldt-Jakob disease
- Crohn's disease
- Drug-induced liver injury
- Eosinophil count
- Essential tremor
- Exfoliation glaucoma
- F cell distribution
- Fasting glucose
- Folate pathway vitamins
- Freckles and burning
- Gallstones
- Hair color
- Heart rate
- Height
- Hepatitus B
- Hirschsprung's disease
- HDL cholesterol
- Idiopathic pulmonary fibrosis
- Inflammatory bowel disease
- Intracranial aneurysm
- Iris color
- Ischemic stroke
- Juvenile idiopathic arthritis
- Kidney stones
- LDL cholesterol
- Liver enzymes
- Lung cancer
- Malaria
- Male pattern baldness
- MCP-1
- Mean platelet volume
- Melanoma
- Menarche & menopause
- Multiple sclerosis
- Myeloproliferative neoplasms
- Narcolepsy
- Neuroblastoma
- Nicotine dependence
- Obesity
- Otosclerosis
- Other metabolic traits
- Panic disorder
- Peripheral arterial disease
- Plasma LP (a) levels
- Primary biliary cirrhosis
- Prostate cancer
- Protein levels
- Pulmonary funct. COPD
- QT interval prolongation
- Psoriasis
- Recombination rate
- Red vs.non-red hair
- Renal function
- Restless legs syndrome
- Rheumatoid arthritis
- Serum bilirubin
- Serum IgE levels
- Serum markers of iron status
- Serum metabolites
- Serum urate
- Skin pigmentation by reflectance spectroscopy
- Soluble ICAM-1
- Statin-induced myopathy
- Stroke
- Systemic lupus erythematosus
- Systemic lupus erythematosus in women
- Testicular germ cell tumor
- Thyroid cancer
- Total cholesterol
- Triglycerides
- Type 1 diabe
- Type 2 dia
- Venous Thromboembolism
- Vitamin B12 levels
- Warfarin dose
- Weight
- YKL-40 levels

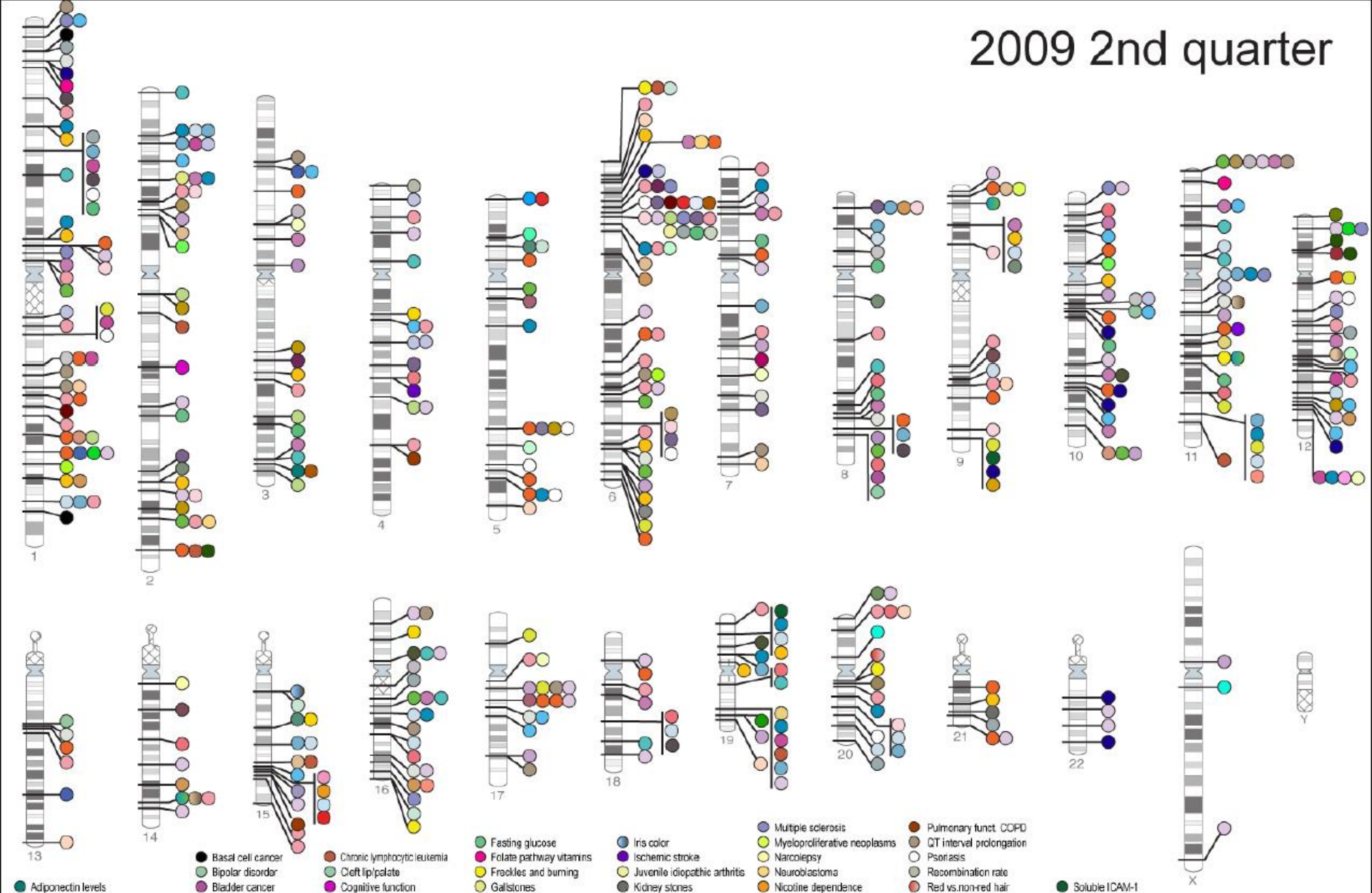
2007 2nd quarter

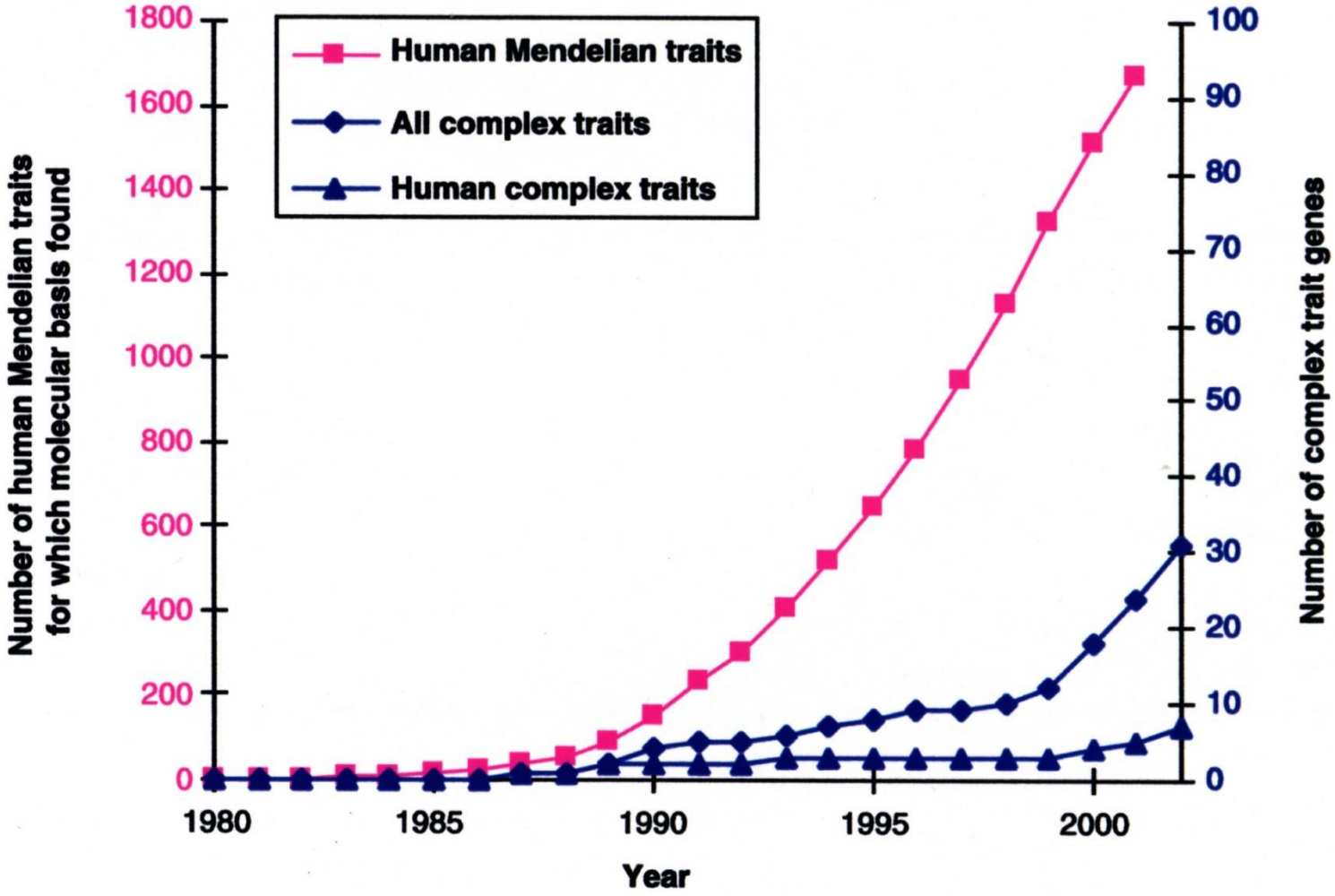


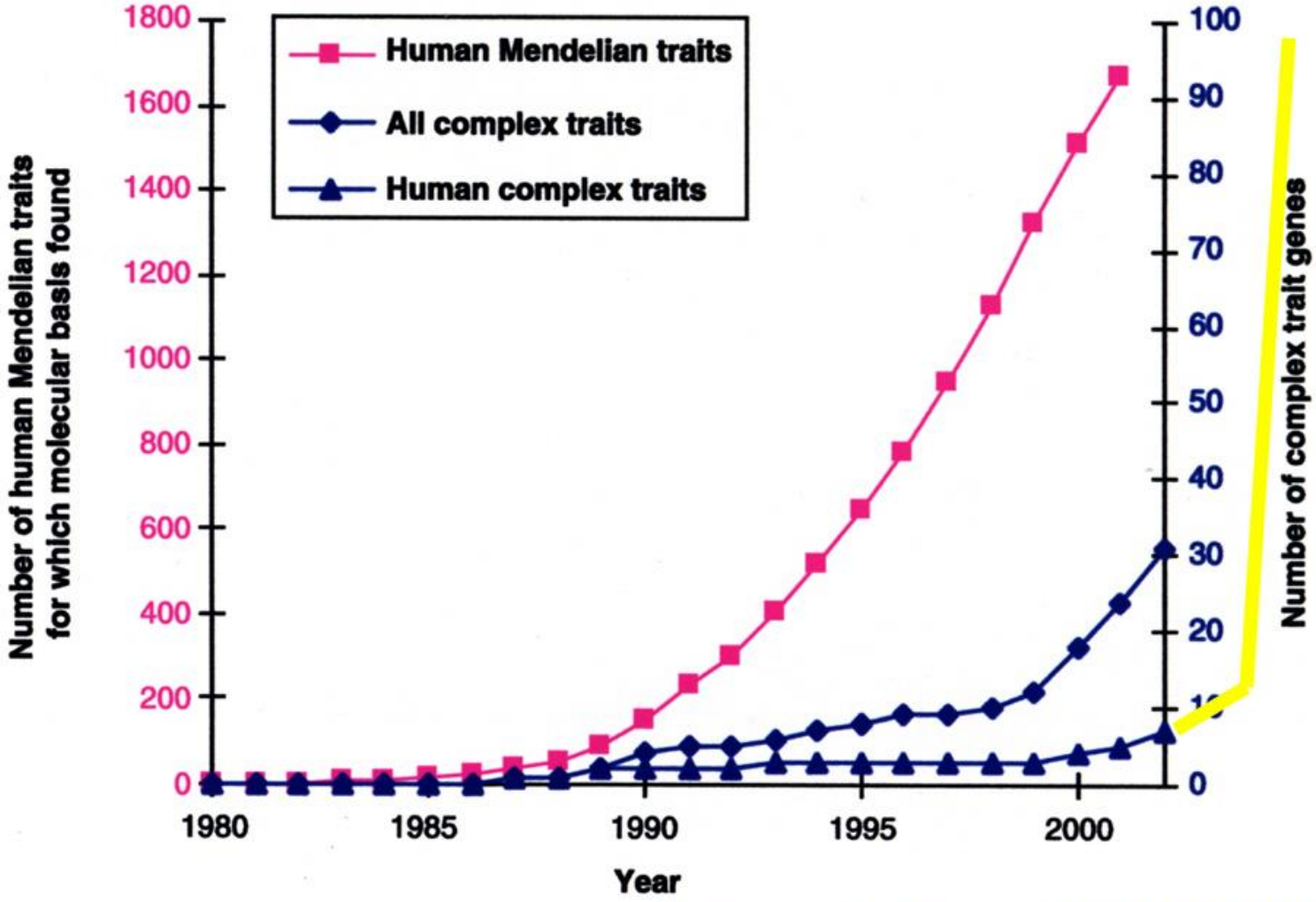
2008 2nd quarter



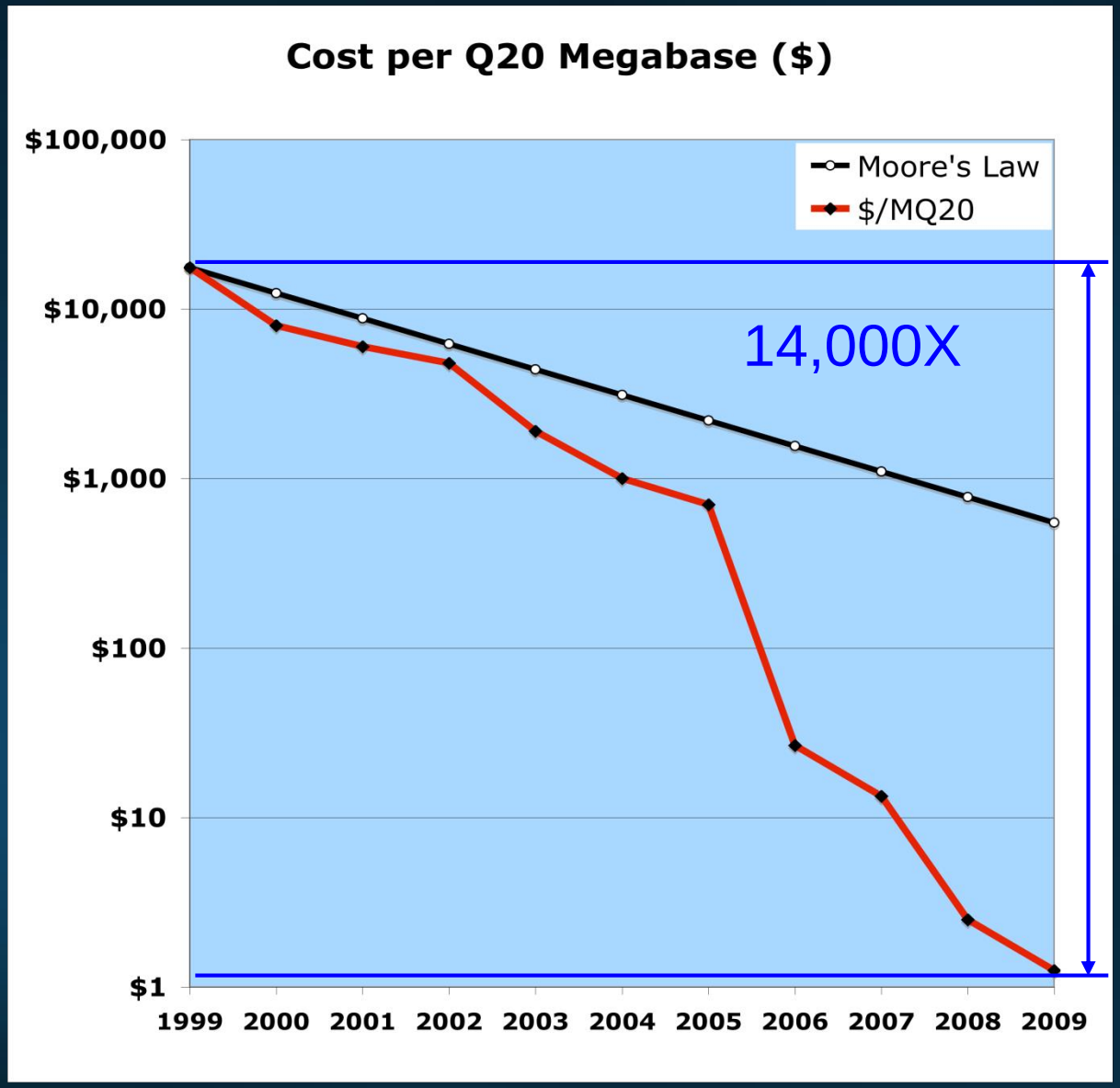
2009 2nd quarter




Human Mendelian traits
All complex traits
Human complex traits
Number of human Mendelian traits for which molecular basis found
Number of complex trait genes
Year
1800
1600
1400
1200
1000
800
600
400
200
0
100
90
80
70
60
50
40
30
20
10
0
1980
1985
1990
1995
2000



Glazier et al., Science 298:2345-9, 2002


Cost per Q20 Megabase ($)
$100,000
$10,000
$1,000
$100
$10
$1
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Moore's Law
$/MQ20
14,000X

An information explosion…

- Lots of data in genome
- More data in when we attempt to
  - discern structure of data
  - relate to transciptomics, proteomics
  - relate to structure, physiology
  - relate to disease
  - relate to variation
- Automated discovery, experiments
- Biomedical knowledge (coming)
- Clinical knowledge (coming)

[Some] Research Projects

- **The Human Genome Project** -- old news, 6 years ago

- **International HapMap Project** -- www.hapmap.org
- **The 1000 Genomes Project** – www.1000.genomes.org
- Encyclopedia of DNA Elements (**ENCODE**) Project
- **The Cancer Genome Atlas** (TCGA)

- **Human Microbiome Project** (HMP) – www.hmpdacc.org

- The **eMERGE** (Electronic Medical Records and Genomics) Network

Common Features of Projects

- High throughput
- Use of technology, in particular
  - Automation (Robotics, AI)
  - Databases
  - Visualization, simulation/computational models
  - Groupware: Coordination and communication
- Public domain tools
- Open sharing of data

Some Challenges

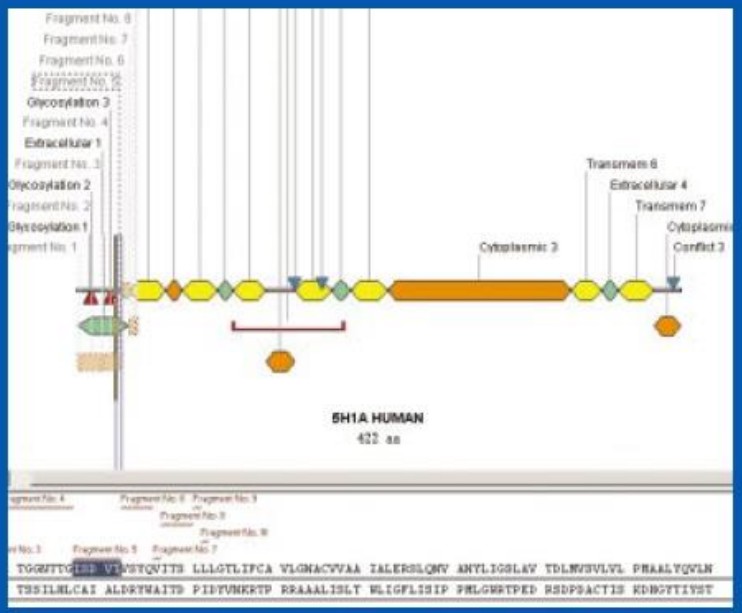
- Volume of data is staggering
  - How to store and collect sequence information?
  - RDBMSs don't handle sequence data well
  - Better handled by Object Oriented DBM
- How to analyze and display the data
  - Automated algorithms
  - Contextual visualization methods
    - Clusters, profiles, etc
- Sequence data is meaningless without context
  - Not well suited to printed medical record

General Informatics Techniques/Tools in Bioinformatics

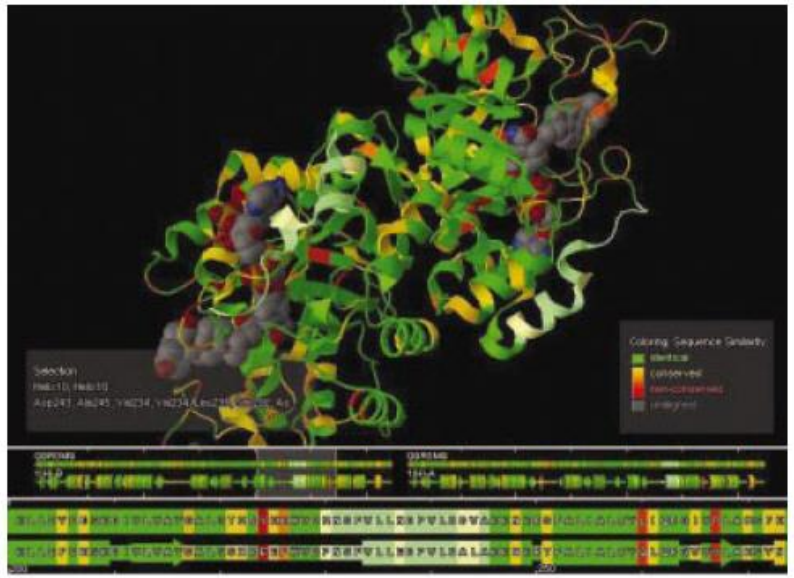
- Storage
  - Databases
    - Building, Querying
    - Complex data
    - Annotations
    - Citations
- Standards
- Interoperability
- Knowledge Management
  - Classification
  - Vocabularies
  - Ontologies
- Communications
- Process Workflow
- Discovery and Analyses
  - Text String Comparison
    - Text search
    - Statistical analysis
  - Finding Patterns
    - AI / Machine Learning
    - Clustering
    - Data mining
  - Geometric
    - Robotics
    - Graphics (Surfaces, Volumes)
    - Comparison and 3D Matching (Vision, Recognition)
  - Physical Simulation
    - Newtonian Mechanics
    - Electrostatics
    - Numerical Algorithms
    - Simulation

Bioinformatics: Tools

- Annotation
- Visualization



InforMax's BioAnnotator uses locally stored databases to find protein motifs.



Structure prediction: modelling a sequence homolog in LION's SRS 3D.

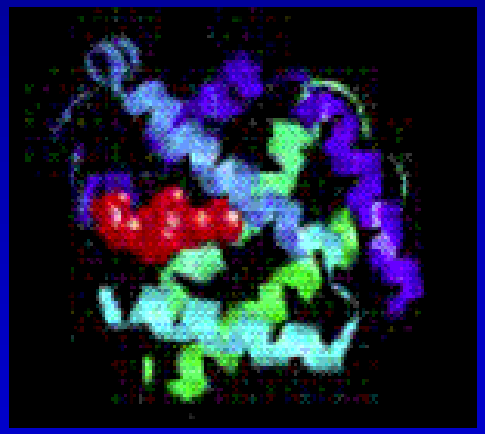
– user friendly, in the public domain, and increasingly integrated

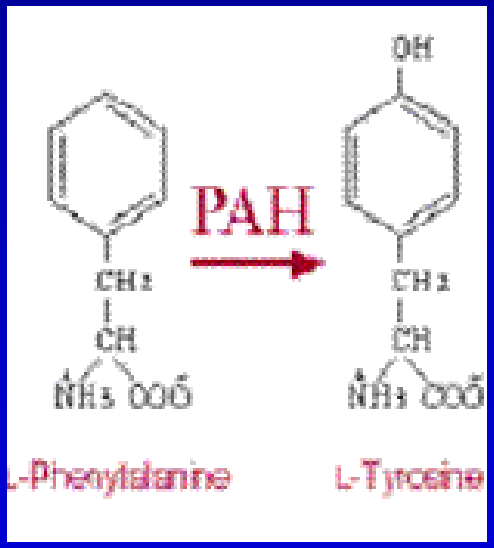
– commercial tools streamline tasks, access proprietary databases

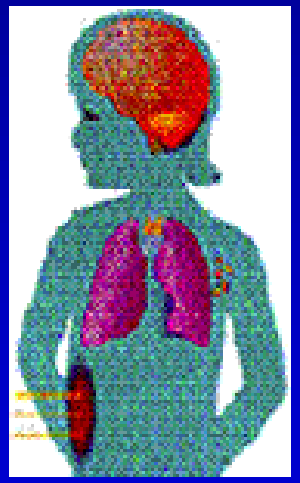
…from Chicurel (2002), p 753-754.

Central Paradigm of Bioinformatics

Genetic Information → Molecular Structure → Biochemical Function → Symptoms

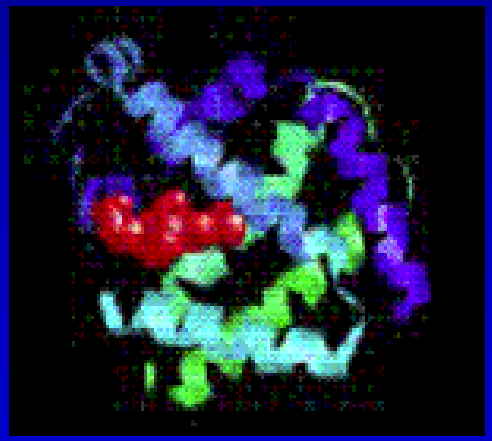


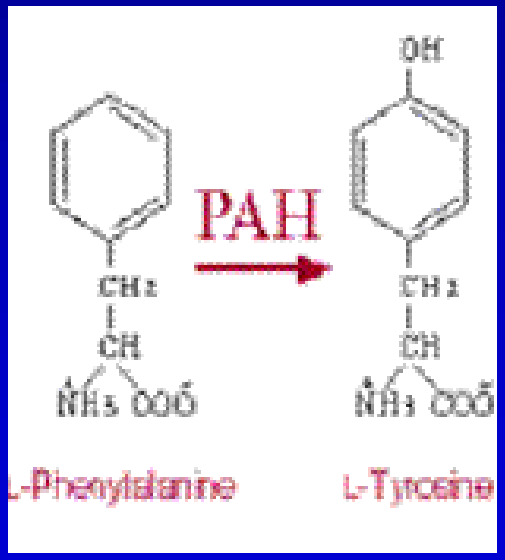


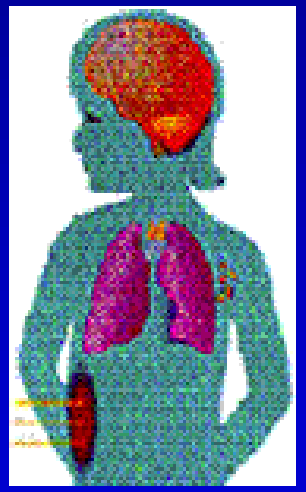


Central Paradigm of Bioinformatics

Genetic Information → Molecular Structure → Biochemical Function → Symptoms







Cosa è la Bioinformatica?

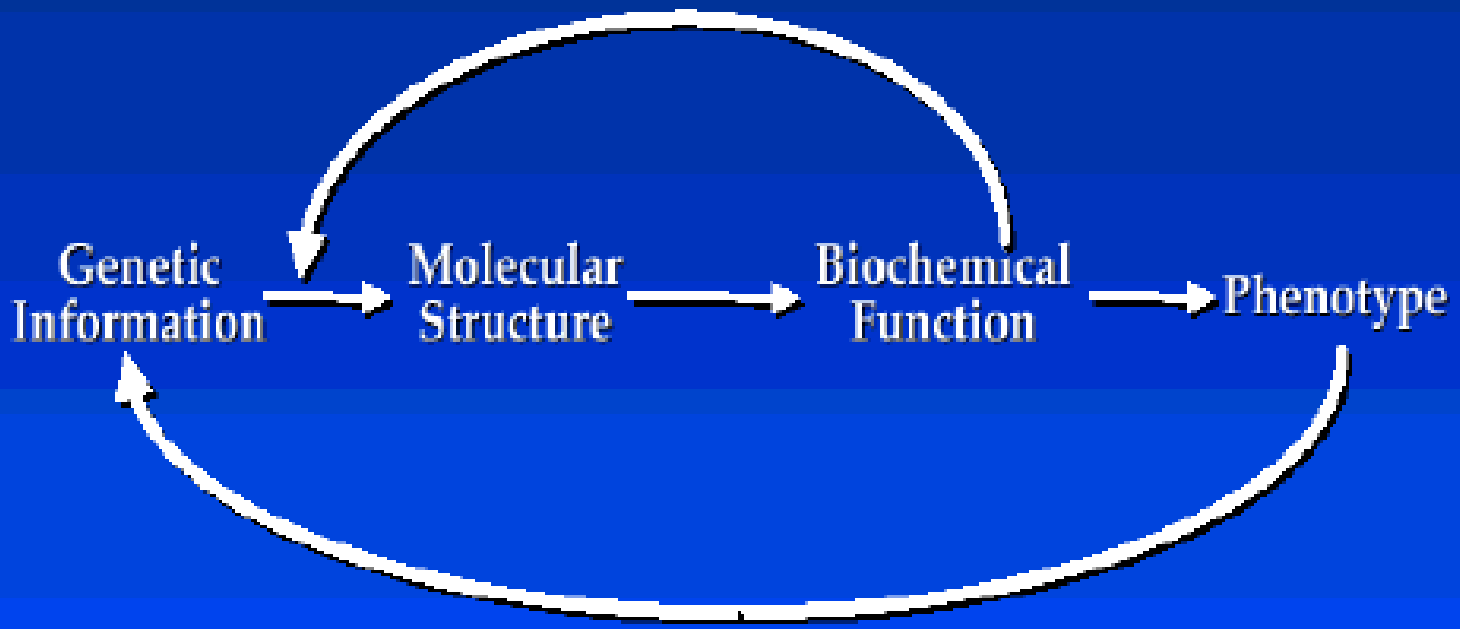
Un campo delle scienze in cui Biologia, Informatica e Biotecnologie confluiscono in un'unica disciplina che permetta:

- **Sviluppo di nuovi algoritmi e funzioni statistiche per individuare relazioni tra informazioni provenienti da diversi gruppi di dati**
- **Analisi e interpretazione di vari tipi di dati**
- **Sviluppo e implementazione di strumenti per accedere e gestire con efficienza diversi tipi di informazioni**

Perché usare la Bioinformatica

Crescita esponenziale di informazioni biologiche (sequenziamento automatico, DNA chips, identificazione proteine, spettrometria di massa ecc.) ha richiesto lo sviluppo di algoritmi che permettessero l'analisi di dati e la loro archiviazione (tools e database)

Central Paradigm of Bioinformatics



Perché è così difficile riuscire ad ottenere informazioni solo dalla sequenza:

L'informazione genetica è ridondante

- **Codice genetico**
- **Sostituzione amminoacidica**
- **Variazioni introni-esoni**
- **Variazione del modulo di lettura**

L'informazione strutturale è ridondante

- **Cambiamenti conformazionali**
- **Differenti strutture possono mostrare similare funzione**
- **Differenti sequenze danno simili strutture**

Geni singoli possono avere diverse funzioni

- **Agire come enzimi metabolici e regolatori**

I geni hanno 1D ma la funzione dipende dalla struttura 3D

NCBI

National Center for Biotechnology Information

National Library of Medicine — National Institutes of Health

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

Search GenBank for Go

NCBI Homepage

SITE MAP

About NCBI
general and contact information

GenBank
sequence submission support and software

Molecular databases
sequences, structures and taxonomy

Literature databases
PubMed, OMIM and PubMed Central NEW

Genomic biology
the human genome, whole genomes and related resources

Tools
for data mining

Research at NCBI

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease.

Draft Human Genome

Explore human genome resources or browse the human genome sequence using the Map Viewer.

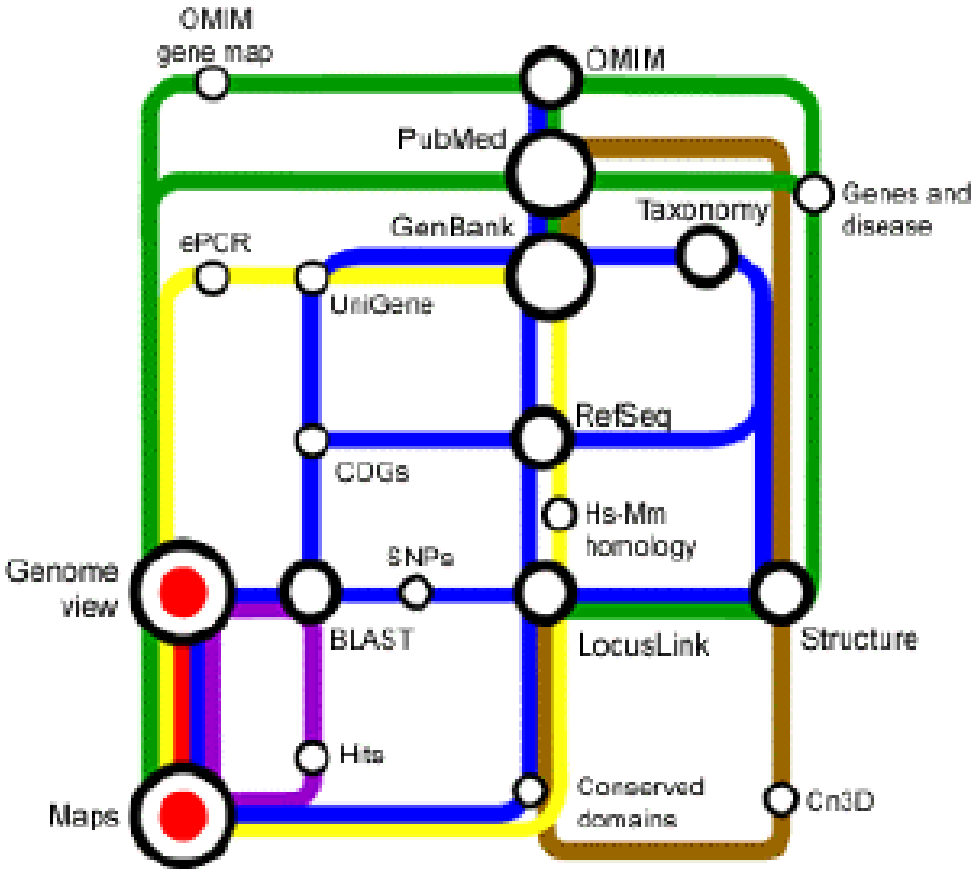
DART: A new tool Archives

Want to locate protein neighbors by domain architecture? Learn about NCBI's new Domain Architecture Retrieval Tool...

http://www.ncbi.nlm.nih.gov/

Hot Spots

- Cancer genome anatomy project
- Clusters of orthologous groups
- Coffee Break
- Electronic PCR
- Gene expression omnibus
- Genes and disease
- Human genome resources
- Human map viewer
- Human/mouse homology maps



Key

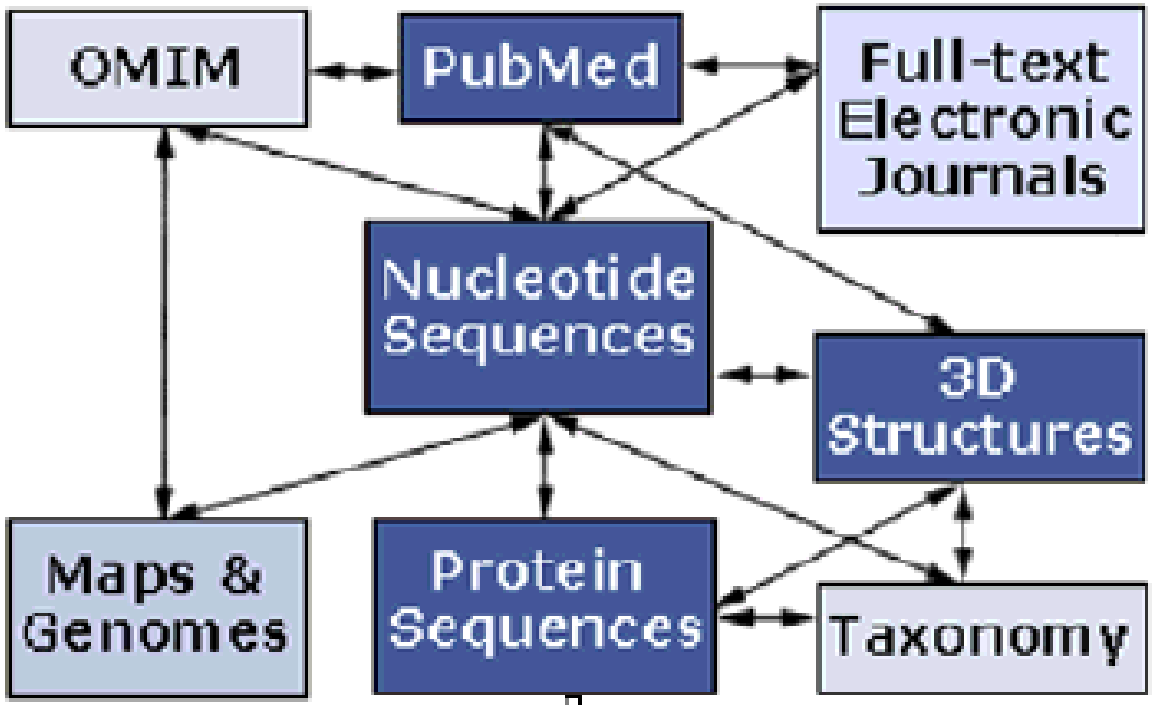
- Literature
- Sequence
- Mapping
- BLAST
- Genome View
- Structure

Back to showcase

http://www.ncbi.nlm.nih.gov/Tour/tour.html

ENTREZ

A search and retrieval system for information integration.



Entrez is a retrival system for searching several linked databases.

It provides access to:

MedLine, abstracts and links to journals

PubMed:	The biomedical literature (PubMed)
Nucleotide:	Sequence databases (GenBank)
Protein:	Sequence databases
Structure:	Three-dimentional macromolecular structures
Genome:	Complete genome assemblies
PopSet:	Population study data sets
Taxonomy:	Organisms in GenBank
OMIM:	Online Mendelian Inheritance in Man

GenBank,EMBL, DDBJ, RefSeq

GenBank, DDBJ, EMBL, PDB, PIR, Swiss- prot,RefSeq

NCBI's MMDB- derived from PDB

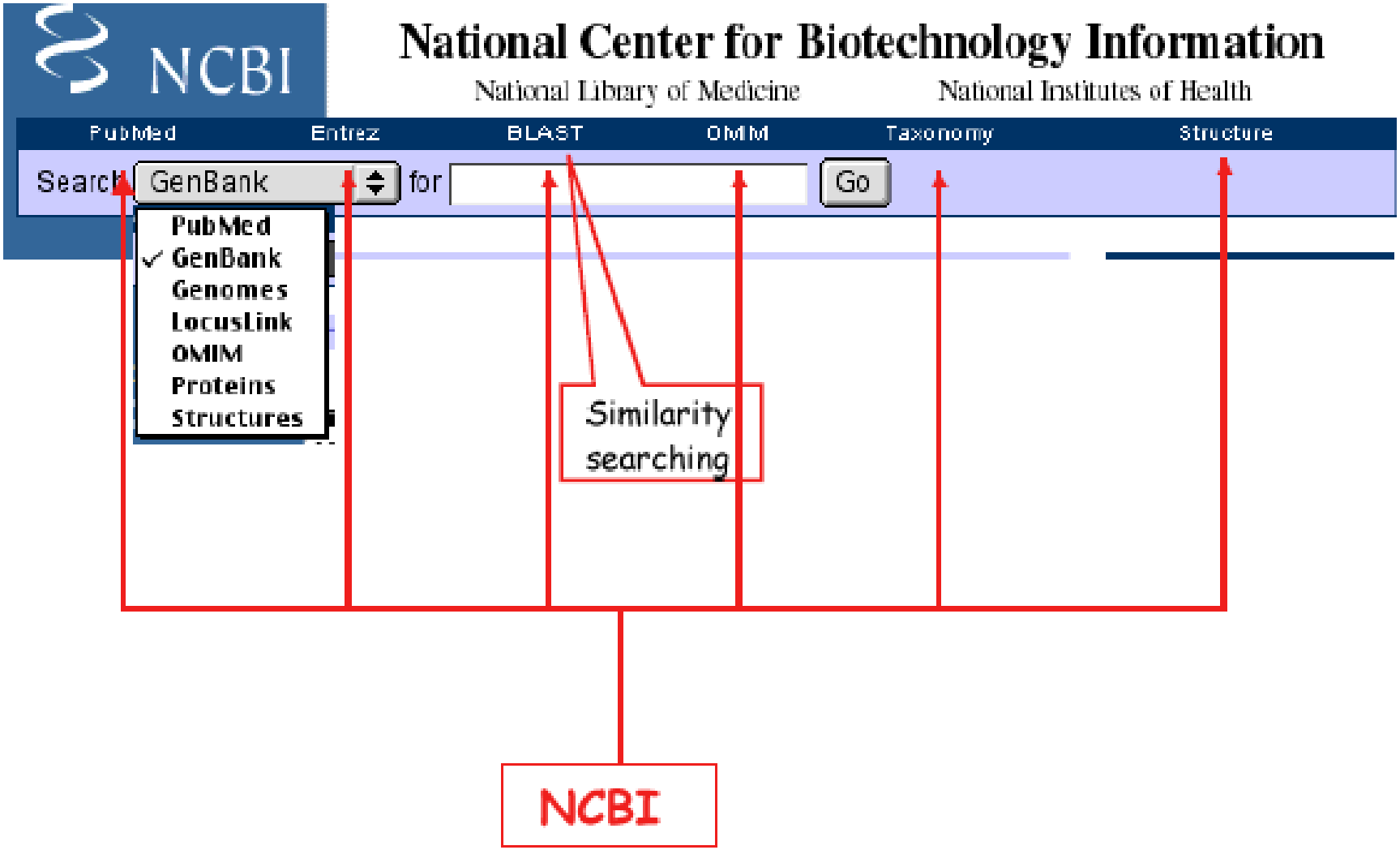
Graphical view, mapping data

Population and phylogenetic studies

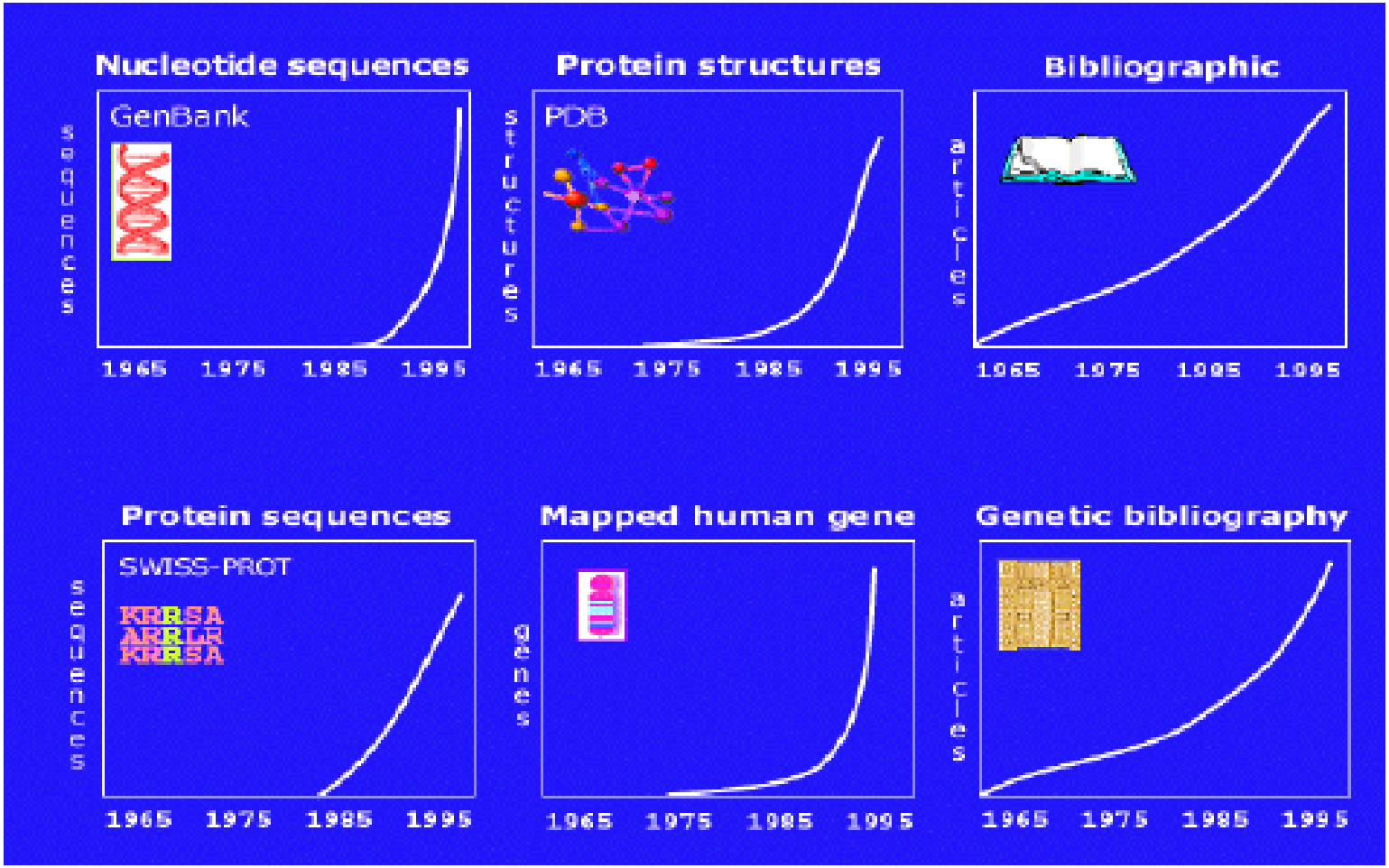
MIM in Entrez

NCBI's Taxonomy, tree structures

http://www.ncbi.nlm.nih.gov/Entrez/


NCBI
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
PubMed
Entrez
BLAST
OMIM
Taxonomy
Structure
Search
GenBank
for
Go
PubMed
GenBank
Genomes
LocusLink
OMIM
Proteins
Structures
Similarity searching
NCBI

Biological Information



- Exponential growth of biological information: growth of sequences, structures, and literature.

- Efficient storage and management tools were most important.

NCBI bioinformatics tools - 1-

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

Search GenBank for Go

BLAST

COGs

Map Viewer

LocusLink

UniGene

ORF finder

Electronic PCR

BLAST **The Basic Local Alignment Search Tool** (BLAST), for comparing gene and protein sequences against others in public databases, now comes in several flavors including PSI-BLAST, PHI-BLAST, and BLAST 2 sequences. Specialized BLASTs are also available for human, microbial, and malaria genomes, as well as for vector contamination, immunoglubulins, and tentative human consensus sequences.

Clusters of Orthologous Groups (COGs) currently covers 21 complete genomes from 17 major phylogenetic lineages. A COG is a cluster of very similar proteins found in at least three species. The presence or absence of a protein in different genomes can tell us about the evolution of the organisms, as well as point to new drug targets.

Map Viewer shows integrated views of chromosome maps currently for human, mouse, and Drosophila. Used to view the NCBI assembly of the complete human genome, Map Viewer is a valuable tool for the identification and localization of genes that contribute to human disease.

LocusLink combines descriptive and sequence information on human genes through a single query interface. LocusLink covers information on official nomenclature, aliases, sequence accession numbers, phenotypes, EC numbers, OMIM numbers, UniGene clusters, map information, and relevant web sites.

NCBI bioinformatics tools -2-

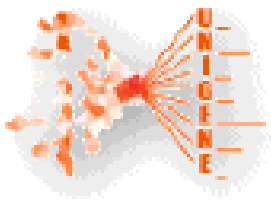
VAST search

CCAP

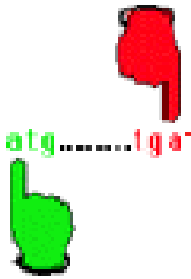
Human-Mouse

VecScreen

CGAP

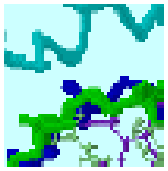


A **UniGene** cluster is a non-redundant set of sequences that represents a unique human, mouse, or rat gene. Well-characterized genes, as well as thousands of expressed sequence tag (EST) sequences have been included. Each cluster record also contains information such as the tissue types in which the gene has been expressed and map location. UniGene can assist in gene discovery, gene mapping projects, and large-scale expression analysis.

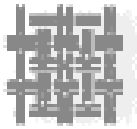


ORF finder identifies all possible ORFs in a DNA sequence by locating the standard and alternative stop and start codons. The deduced amino acid sequences can then be used to BLAST against GenBank. ORF finder is also packaged in the sequence submission software Sequin.

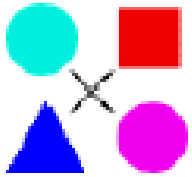
Electronic PCR allows you to search your DNA sequence for sequence tagged sites (STSs), which have been used as landmarks in various types of genomic maps. It compares the query sequence against data in NCBI's UniSTS, a unified, non-redundant view of STSs from a wide range of sources.



VAST search is a structure-structure similarity search service. It compares 3D coordinates of a newly determined protein structure to those in the MMDB/PDB database. VAST Search computes a list of similar structures that can be browsed interactively, using molecular graphics to view superimpositions and alignments.



The Cancer Chromosome Aberration Project (CCAP) compiles information on the distinct chromosome aberrations that are associated with different cancers. The identification of chromosomal abnormalities by clinicians can enable the diagnosis of, classification of, and treatment selection for a given cancer.



The **Human-Mouse Homology Maps** compare genes in homologous segments of DNA from human and mouse sources, sorted by position in each genome. A total of 1793 loci are presented, most of which are genes. This map should be interpreted as a reflection of probable, not confirmed, homology relationships due to the lack of further information available for about half the loci.

VecScreen is a tool for identifying segments of a nucleic acid sequence that may be of vector, linker or adapter origin prior to sequence analysis or submission. VecScreen was developed to combat the problem of vector contamination in public sequence databases.

The Cancer Genome Anatomy Project (CGAP) aims to decipher the molecular anatomy of cancer cells. CGAP develops profiles of cancer cells by comparing gene expression in normal, precancerous, and malignant cells from a wide variety of tissues.

Introduzione

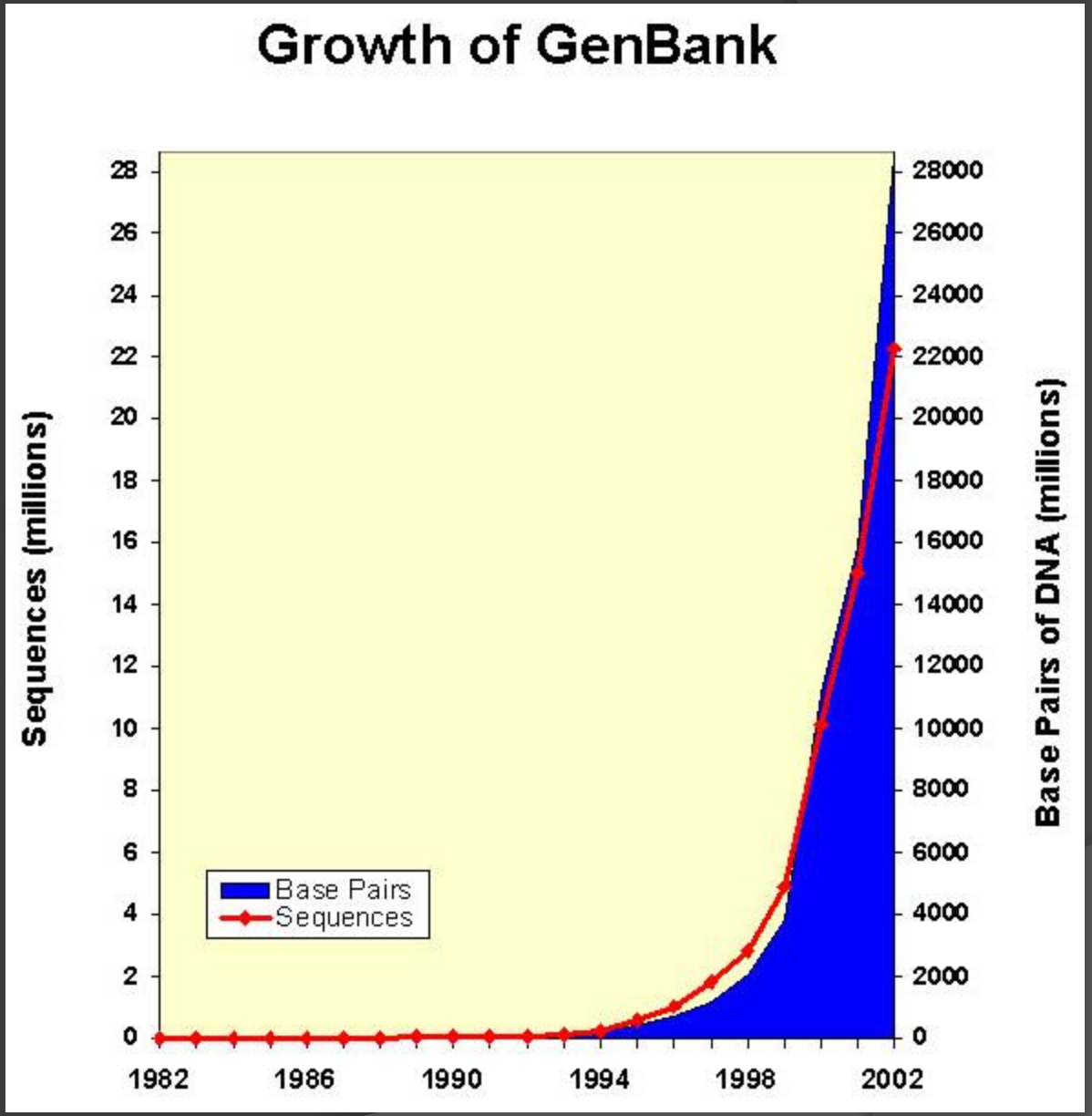
- Esistono diverse banche dati in cui sono depositate sequenze biologiche
  - di nucleotidi (DNA)
    - ✓ GenBank (National Institute of Health, Bethesda, MD)
    - ✓ EMBL (European Bioinformatics Institute, Hinxton, UK)
    - ✓ DNA Database of Japan (Mishima, Japan)
  - di amminoacidi (proteine)
    - ✓ PIR (National Biomedical Research Foundation, Washington, DC)
    - ✓ SwissProt/TrEMBL (Università di Ginevra, CH)
    - ✓ PRFDB (Protein Research Foundation, Osaka, Japan)
- Esiste la necessità di evidenziare, tramite confronto, la similarità tra una sequenza studiata e quelle contenute in una banca dati

BANCHE DATI DI SEQUENZE GENOMICHE

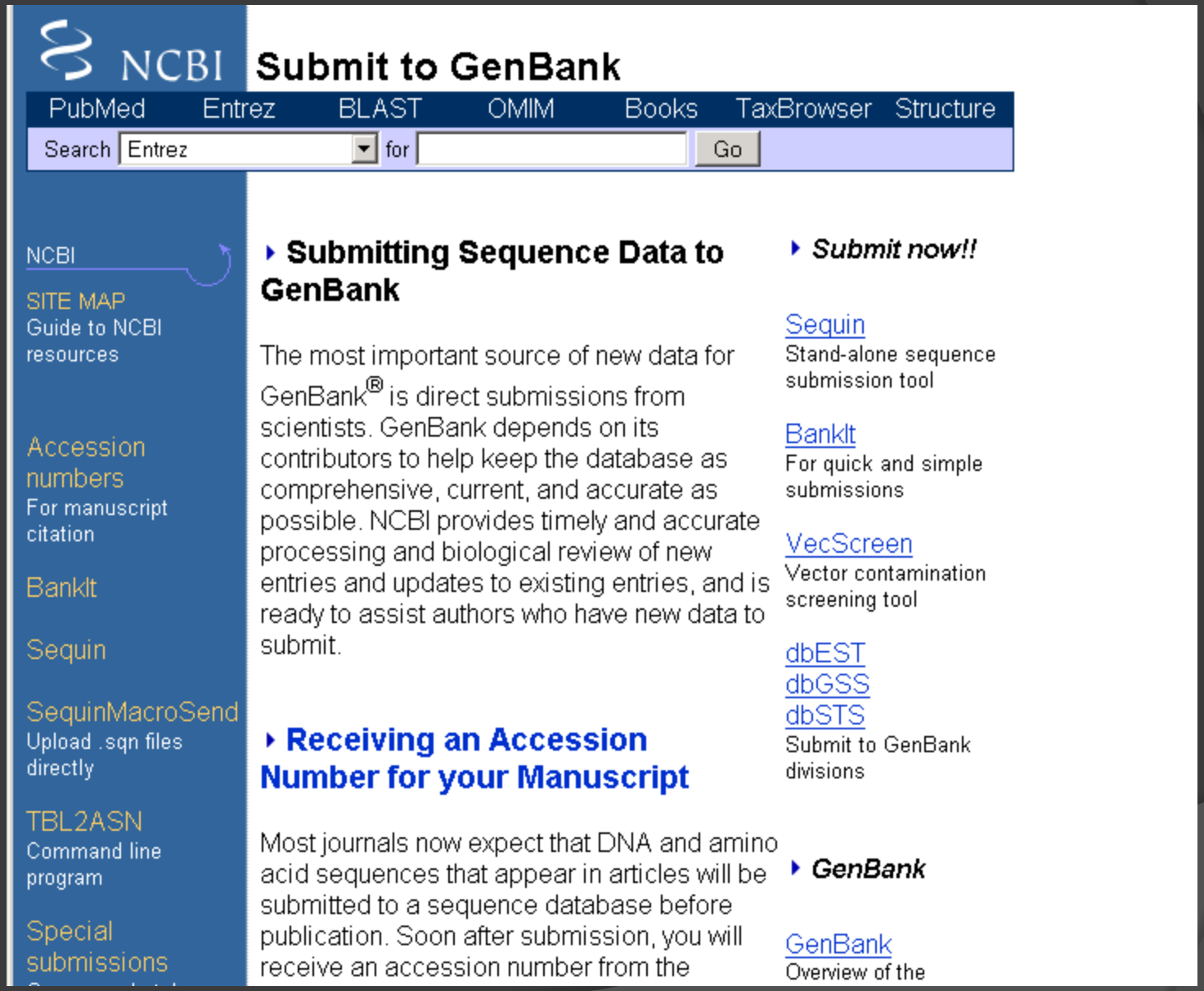
GenBank deriva dalla collaborazione di diversi database di sequenze tra cui EMBL e DDBJ

What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (Nucleic Acids Research 2002 Jan 1;30(1):17-20). There are approximately 22,617,000,000 bases in 18,197,000 sequence records as of August 2002 (see GenBank growth statistics). As an example, you may view the record for a Saccharomyces cerevisiae gene. The complete release notes for the current version of GenBank are available. A new release is made every two months. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.



Genetic Sequence Data Bank February 15 2003 NCBI-GenBank Flat File Release 134.00: 23035823 loci, 29358082791 bases, from 23035823 reported sequences


NCBI
Submit to GenBank
PubMed Entrez BLAST OMIM Books TaxBrowser Structure
Search Entrez for Go
NCBI
SITE MAP
Guide to NCBI resources
Accession numbers
For manuscript citation
BankIt
Sequin
SequinMacroSend
Upload .sqn files directly
TBL2ASN
Command line program
Special submissions
Submitting Sequence Data to GenBank
The most important source of new data for GenBank® is direct submissions from scientists. GenBank depends on its contributors to help keep the database as comprehensive, current, and accurate as possible. NCBI provides timely and accurate processing and biological review of new entries and updates to existing entries, and is ready to assist authors who have new data to submit.
Receiving an Accession Number for your Manuscript
Most journals now expect that DNA and amino acid sequences that appear in articles will be submitted to a sequence database before publication. Soon after submission, you will receive an accession number from the
Submit now!!
Sequin
Stand-alone sequence submission tool
BankIt
For quick and simple submissions
VecScreen
Vector contamination screening tool
dbEST
dbGSS
dbSTS
Submit to GenBank divisions
GenBank
GenBank
Overview of the

BANCHE DATI DI SEQUENZE PROTEICHE

SWISS-PROT

Swiss-Prot is a curated protein sequence database which strives to provide a high level of annotations (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases

TrEMBL

The TrEMBL database contains the translations of all coding sequences (CDS) present in the EMBL Nucleotide Sequence Database, which are not yet integrated into Swiss-Prot.

SP-TrEMBL (Swiss-Prot TrEMBL) Contains the entries which should eventually be incorporated into Swiss-Prot and can be considered as a preliminary section of Swiss-Prot as all SP-TrEMBL entries have been assigned Swiss-Prot accession numbers.

REM-TrEMBL (REMaining TrEMBL) Contains the entries that we do not want to include in Swiss-Prot. REM-TrEMBL entries have no accession numbers.

Site Map | Search ExPASy | Contact us

Search Swiss-Prot/TrEMBL for Go Clear

ExPASy Proteomics Server

The ExPASy (**Ex**pert **P**rotein **A**nalysis **Sy**stem) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE (Disclaimer / References).

[Announcements] [Job opening] [Mirror Sites]

Databases

- Swiss-Prot and TrEMBL - Protein knowledgebase
- PROSITE - Protein families and domains
- SWISS-2DPAGE - Two-dimensional polyacrylamide gel electrophoresis
- ENZYME - Enzyme nomenclature
- SWISS-3DIMAGE - 3D images of proteins and other biological macromolecules
- SWISS-MODEL Repository - Automatically generated protein models
- GermOnLine - Knowledgebase on germ cell differentiation
- Ashbya Genome Database
- Links to many other molecular biology databases

Tools and software packages

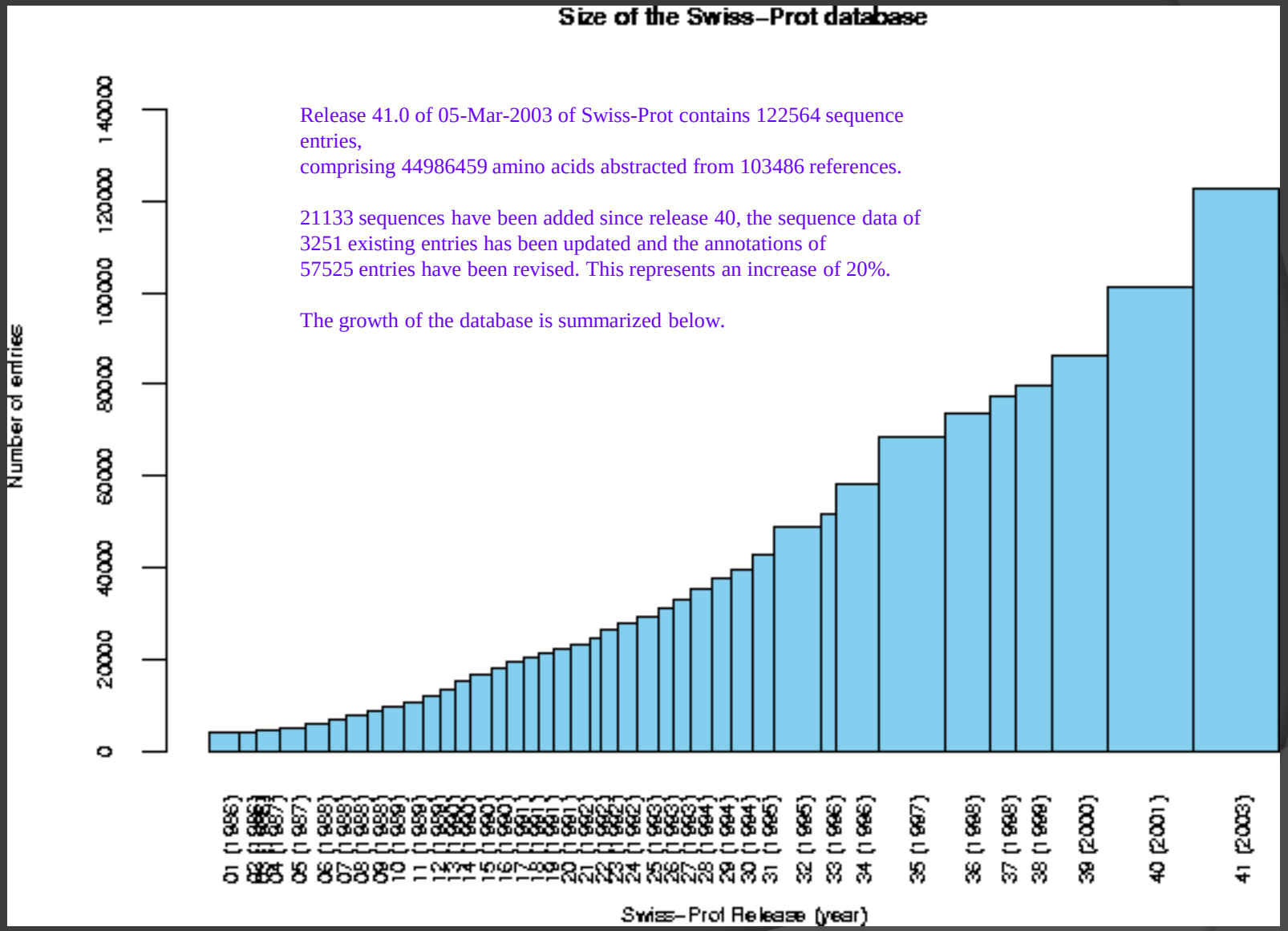
- Proteomics and sequence analysis tools
  - Proteomics [PeptIdent, PeptideMass, ...]
  - DNA -> Protein [Translate]
  - Similarity searches [BLAST]
  - Pattern and profile searches [ScanProsite]
  - Post-translational modification and topology prediction
  - Primary structure analysis [ProtParam, pI/MW, ProtScale]
  - Secondary and tertiary structure prediction [SWISS-MODEL, Swiss-PdbViewer]
  - Alignment [T-COFFEE, SIM]
  - Biological text analysis
- ImageMaster / Melanie - Software for 2-D PAGE analysis
- Roche Applied Science's Biochemical Pathways

Education and services

- The ExPASy FTP server

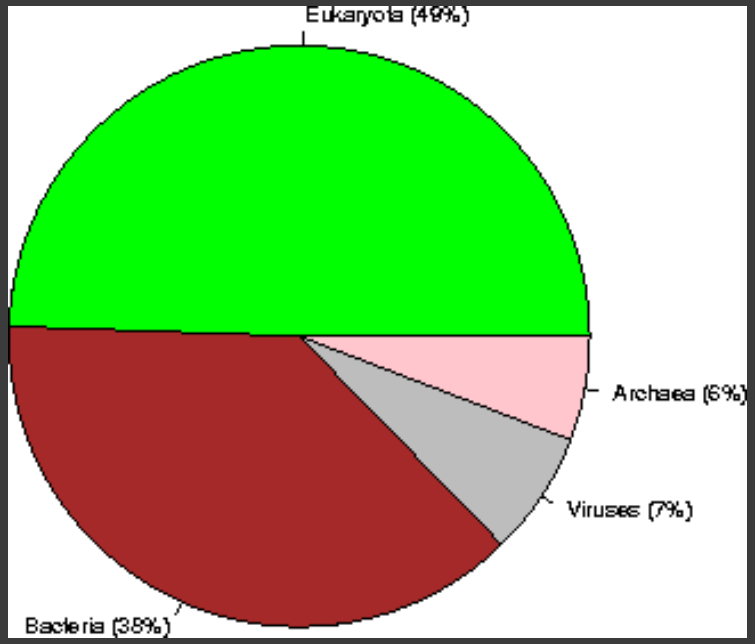
Documentation

- What's New on ExPASy


Size of the Swiss–Prot database
Release 41.0 of 05-Mar-2003 of Swiss-Prot contains 122564 sequence entries,
comprising 44986459 amino acids abstracted from 103486 references.
21133 sequences have been added since release 40, the sequence data of 3251 existing entries has been updated and the annotations of 57525 entries have been revised. This represents an increase of 20%.
The growth of the database is summarized below.
Number of entries
0
20000
40000
60000
80000
100000
120000
140000
32 (1985)
33 (1986)
34 (1986)
35 (1987)
36 (1988)
37 (1988)
38 (1989)
39 (2000)
40 (2001)
41 (2003)
Swiss–Prot Release (year)

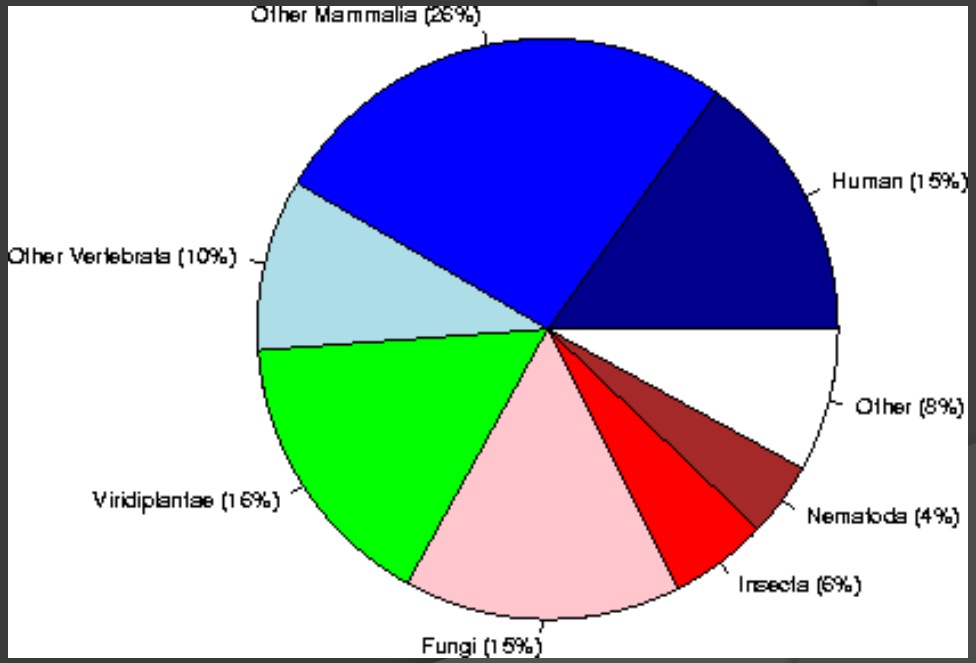
Taxonomic distribution of the sequences

Kingdom	sequences (% of the database)
Archaea	7119 (6%)
Bacteria	46344 (38%)
Eukaryota	60623 (49%)
Viruses	8478 (7%)



Within Eukaryota:

Category	sequences (% of Eukaryota)	(% of the complete datab
Human	9172 (15%)	(7%)
Other Mammalia	16041 (26%)	(13%)
Other Vertebrata	5806 (10%)	(5%)
Viridiplantae	9581 (16%)	(8%)
Fungi	9337 (15%)	(8%)
Insecta	3352 (6%)	(3%)
Nematoda	2504 (4%)	(2%)
Other	4830 (8%)	(4%)



Public Databases:

- NCBI database
- PIR International NBRF USA, MIPS Germania, JIPID Giappone
- SwissProt: Embl (European Bioinformatics Institute) and Swiss Institute
- PDB Databases

Alignment/Similarity Search Tools:

- BLAST
- FASTA
- ClustalW and ClustalX

Pattern/Motif/Properties/Protein Family Finding Tools:

- MotifScan
- Prosite
- Psort
- Protscale
- SOSUI
- Protein strucutre classification CATH and SCOP

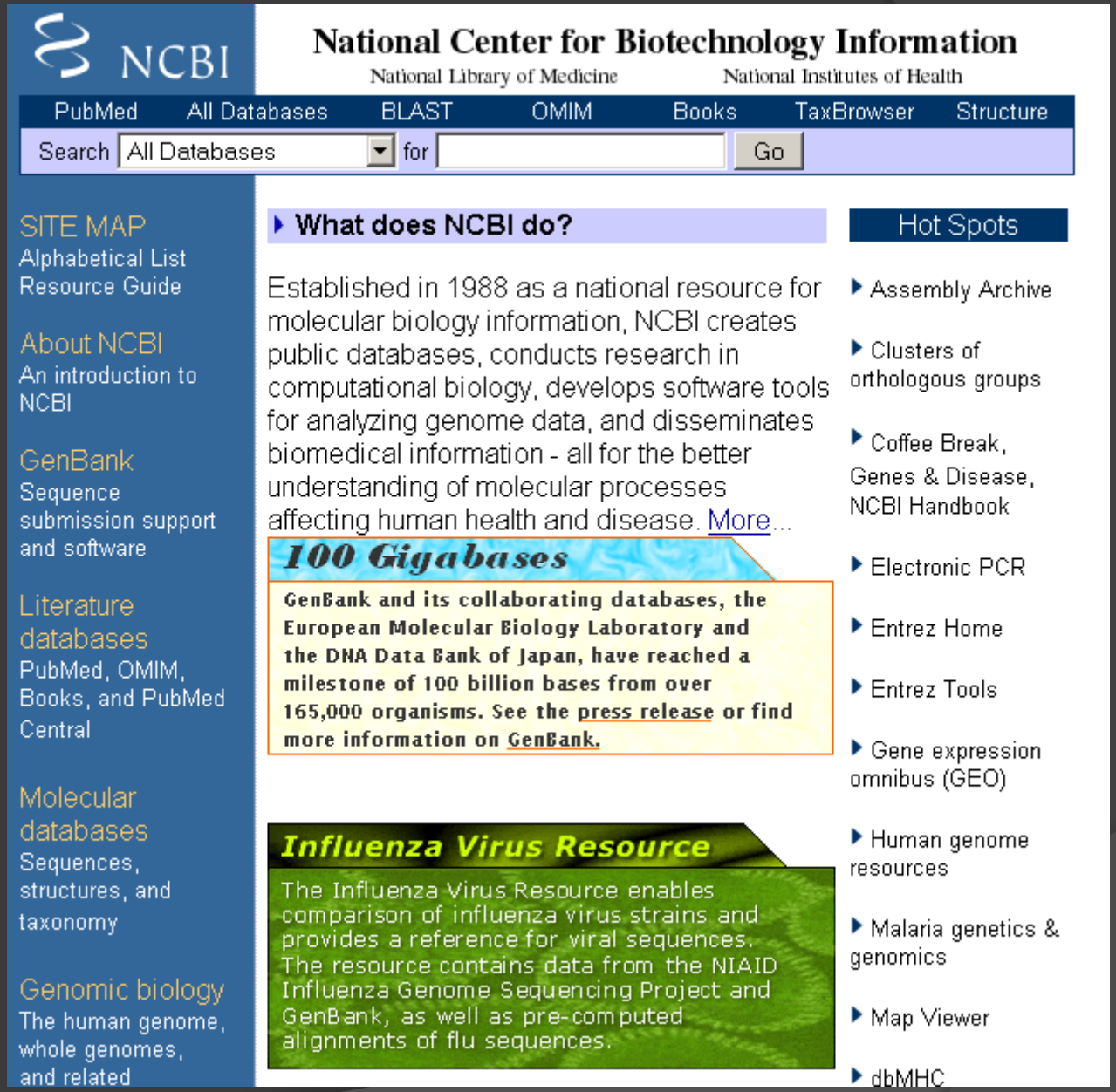
Molecule structure prediction & visulaization Tools:

- Cn3D
- MolMol and DeepViev
- 2D Structure Prediction Tools PHD, PsiPred, GORIV
- SwissModel
- PHYRE

Boutique banche dati specializzate:

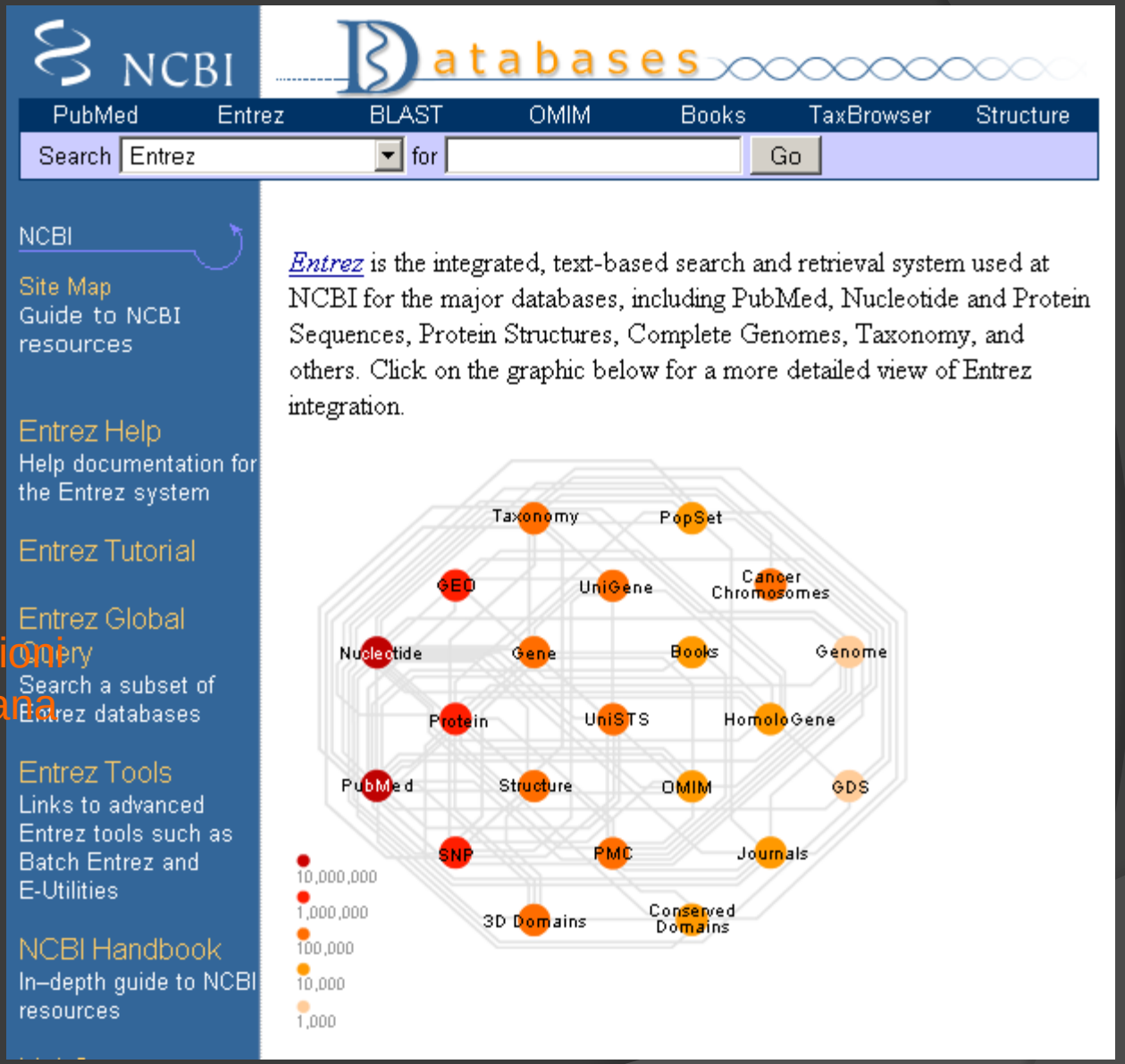
- Proteine chinasi
- Proteasi dell'HIV
- Virus icosaedrici
- Immunologia

Componente della United States National Library of Medicine



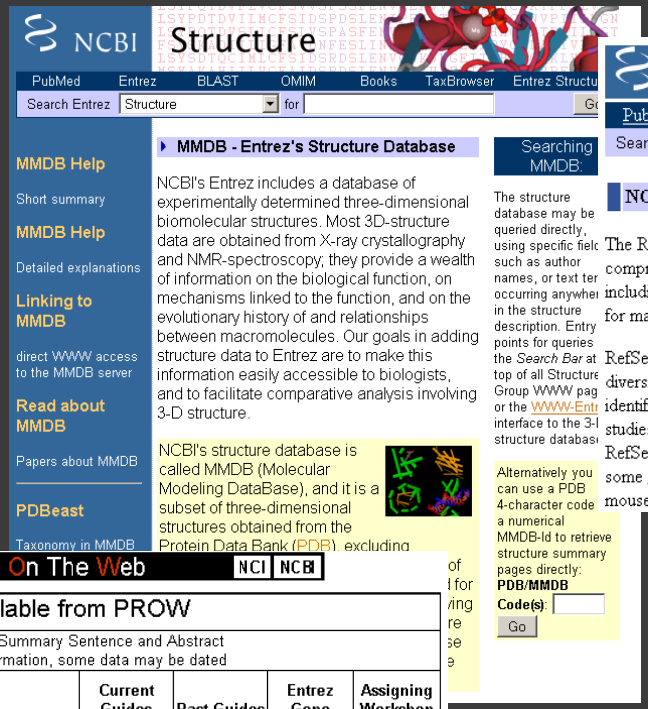
Entrez offre accesso attraverso le seguenti divisioni di banche dati:

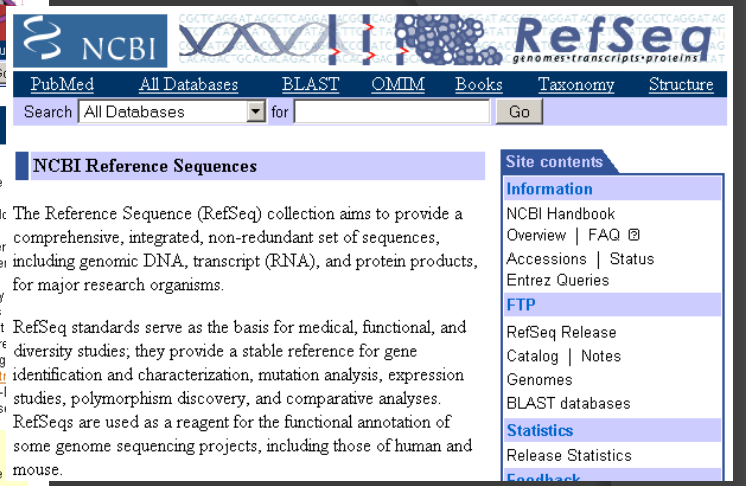
Protein
Peptide
Nucleotide
Structure
Genome
Popset (info su popolazioni)
OMIM (eredità mendeliana nell'uomo)

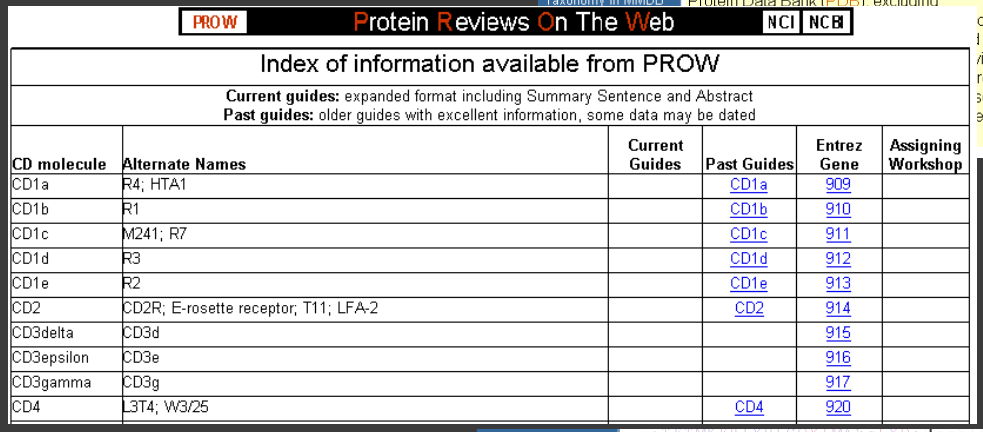


Protein Database

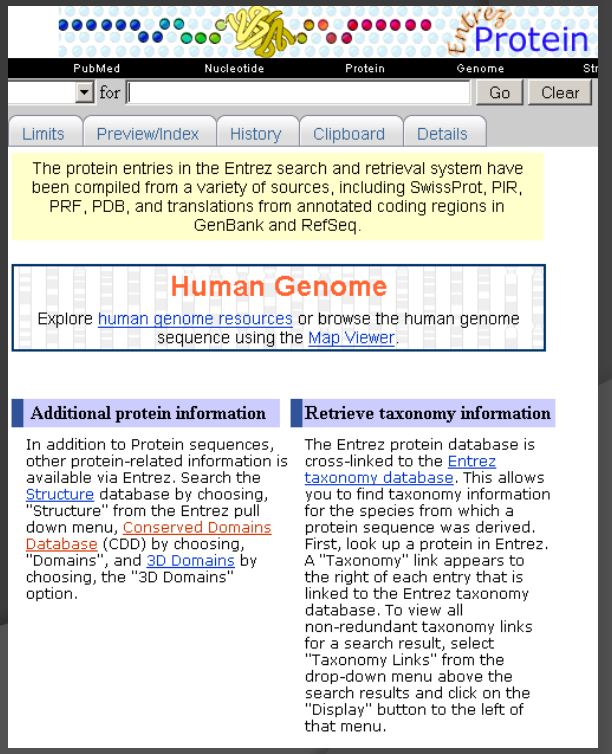
- 3Ddomains
- Protein
- PROW
- RefSeq
- Conserved Domain
- Structure (MMBD)

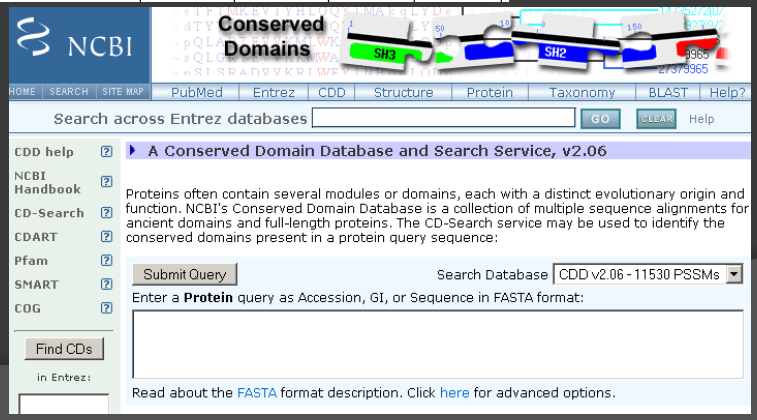


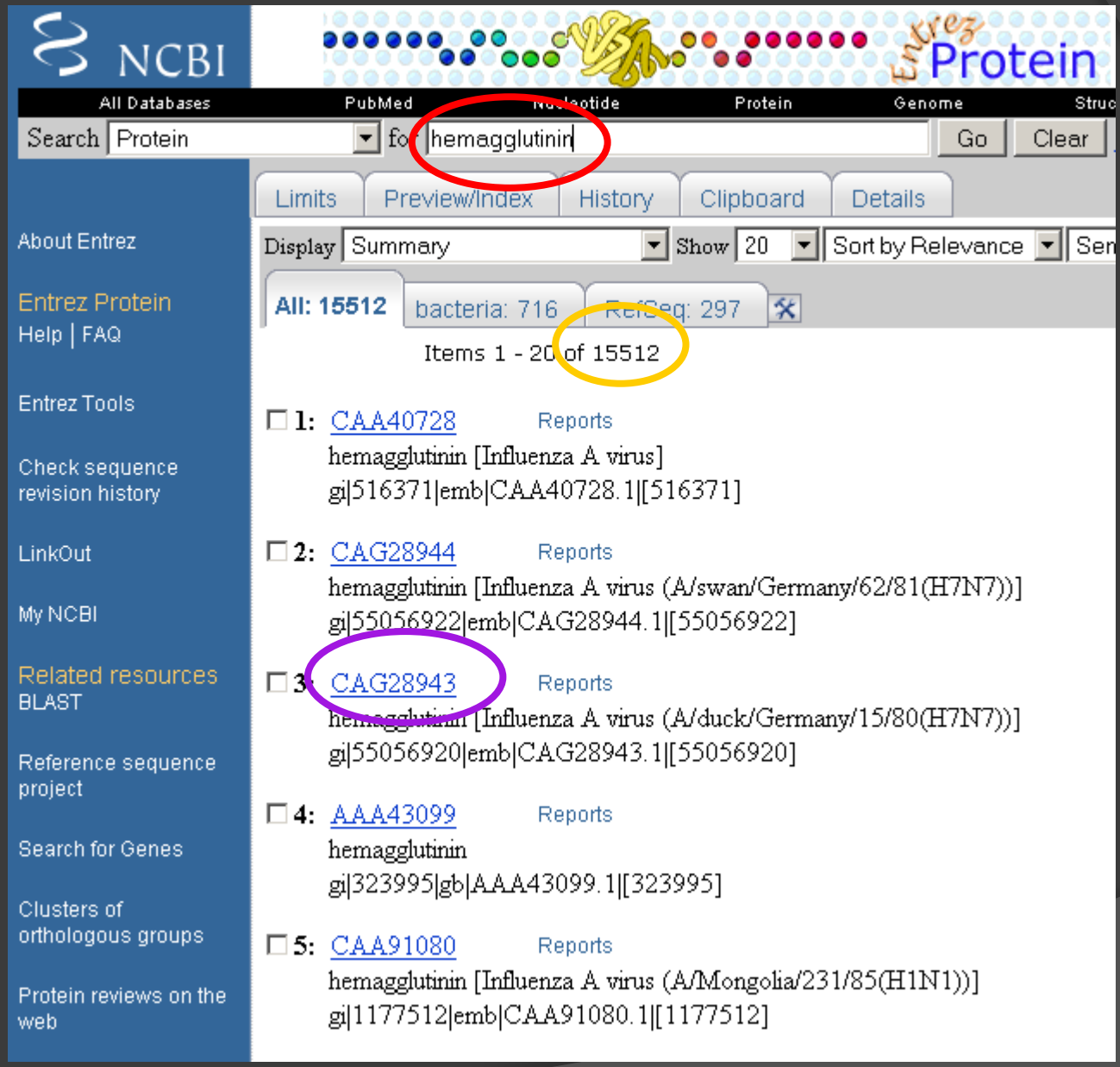




CD molecule	Alternate Names	Current Guides	Past Guides	Entrez Gene	Assigning Workshop
CD1a	R4; HTA1		CD1a	909	
CD1b	R1		CD1b	910	
CD1c	M241; R7		CD1c	911	
CD1d	R3		CD1d	912	
CD1e	R2		CD1e	913	
CD2	CD2R; E-rosette receptor; T11; LFA-2		CD2	914	
CD3delta	CD3d			915	
CD3epsilon	CD3e			916	
CD3gamma	CD3g			917	
CD4	L3T4; W3/25		CD4	920	






NCBI
Entrez Protein
All Databases
PubMed
Protein
Genome
Search Protein for hemagglutinin
Go
Clear
Limits
Preview/Index
History
Clipboard
Details
About Entrez
Entrez Protein
Help | FAQ
Entrez Tools
Check sequence revision history
LinkOut
My NCBI
Related resources
BLAST
Reference sequence project
Search for Genes
Clusters of orthologous groups
Protein reviews on the web
Display Summary
Show 20
Sort by Relevance
All: 15512
bacteria: 716
RefSeq: 297
Items 1 - 20 of 15512
1: CAA40728 Reports
hemagglutinin [Influenza A virus]
gi|516371|emb|CAA40728.1|[516371]
2: CAG28944 Reports
hemagglutinin [Influenza A virus (A/swan/Germany/62/81(H7N7))]
gi|55056922|emb|CAG28944.1|[55056922]
3: CAG28943 Reports
[Influenza A virus (A/duck/Germany/15/80(H7N7))]
gi|55056920|emb|CAG28943.1|[55056920]
4: AAA43099 Reports
hemagglutinin
gi|323995|gb|AAA43099.1|[323995]
5: CAA91080 Reports
hemagglutinin [Influenza A virus (A/Mongolia/231/85(H1N1))]
gi|1177512|emb|CAA91080.1|[1177512]

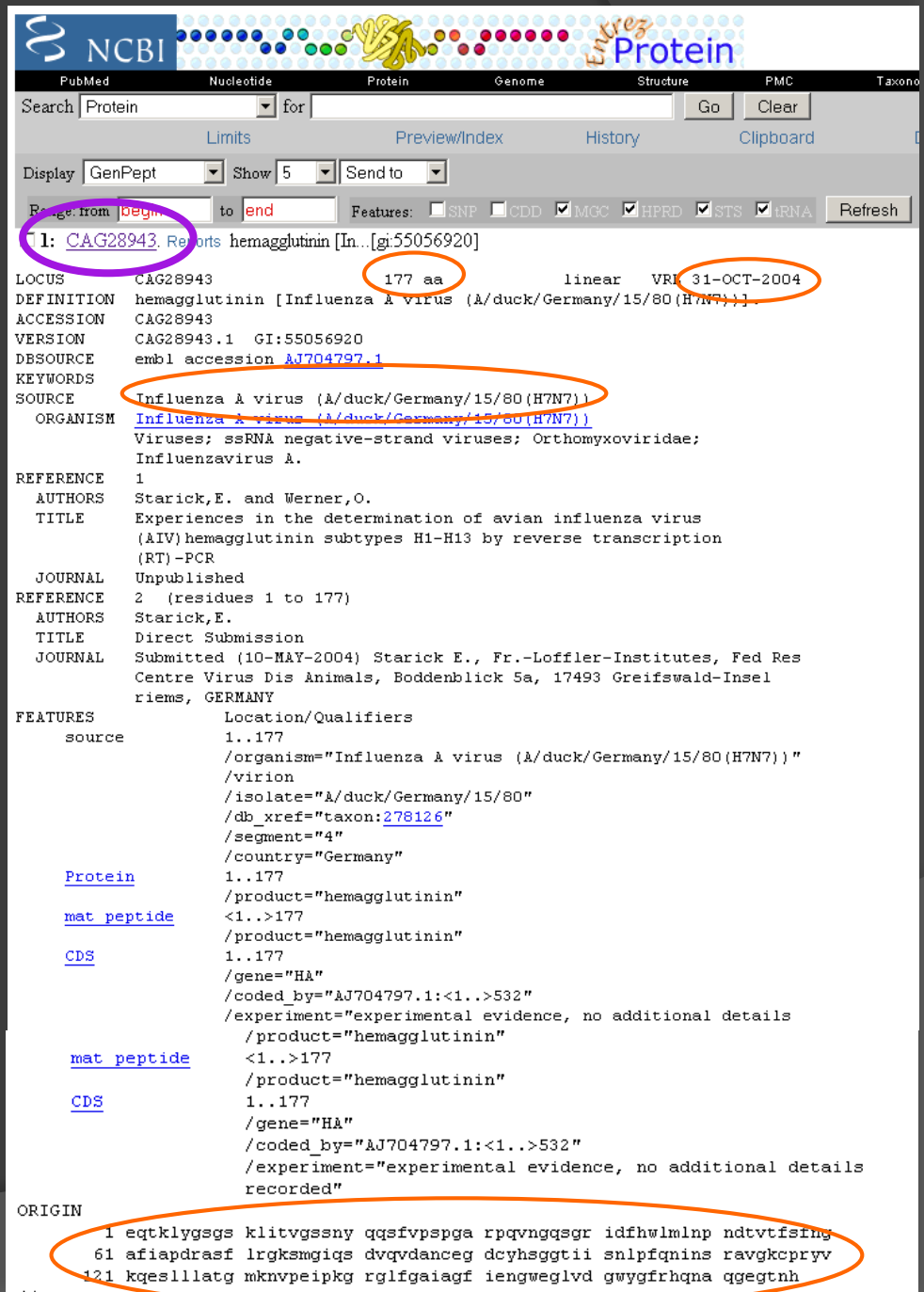
Codice EMB
European Bioinformatics Institute

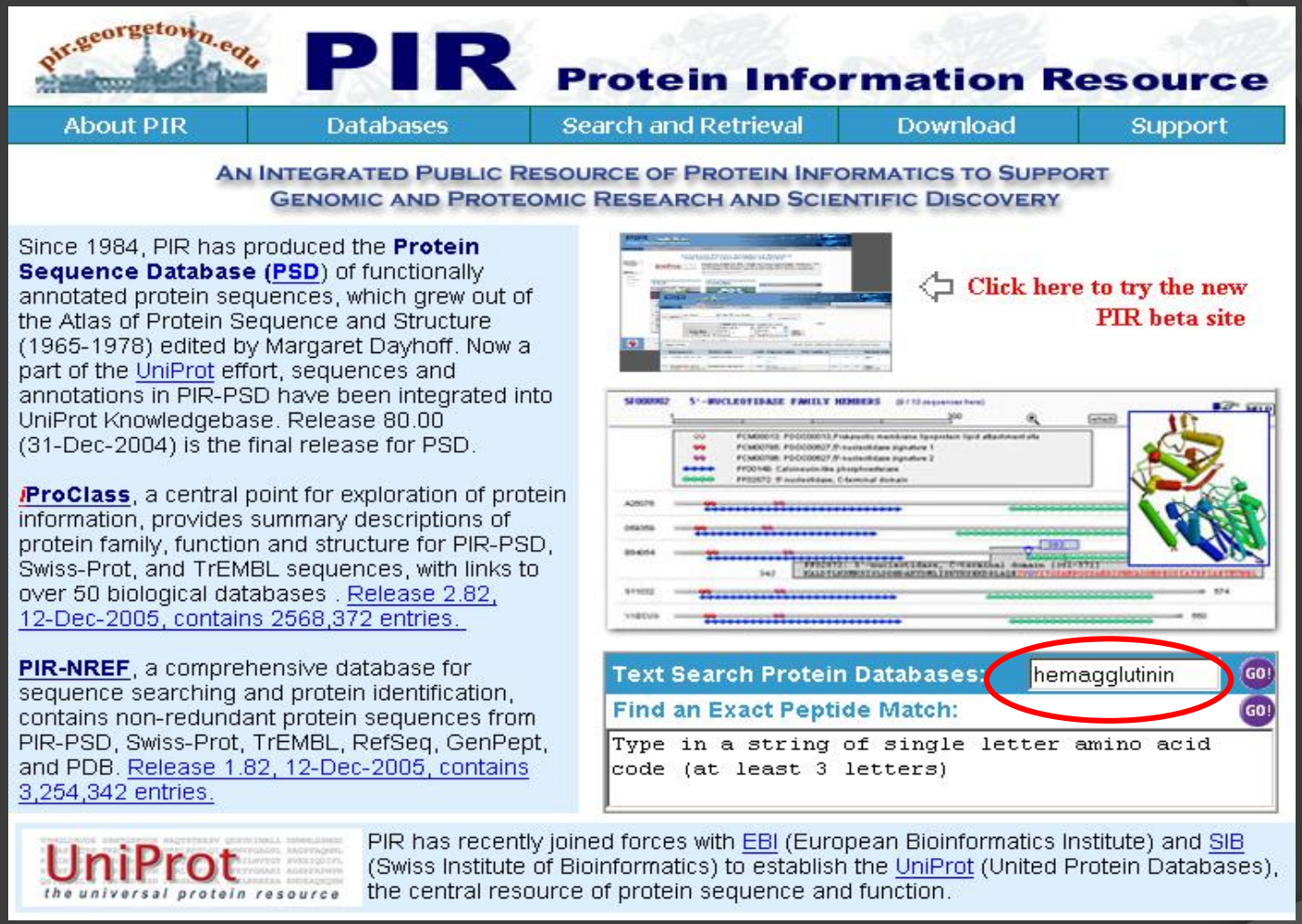
Lunghezza amminoacidica

Data di deposizione

Origine

Sequenza amminoacidica





Combinazione efficace

Di una banca dati accurata

Software per la ricerca di informazioni

Banco di lavoro per lo studio di sequenze

PIR Search Results

pir.georgetown.edu

Site Map Site Search

Text Search Protein Databases: GO!

About PIR | Databases | Search & Retrieval | Download | Support

Thu Jan 26 10:45:30 EST 2006

examples

Search NREF for All Fields hemagglutinin AND All Fields not null AND All Fields AND All Fields

6763 protein sequences in total. page1 (50/page) Next

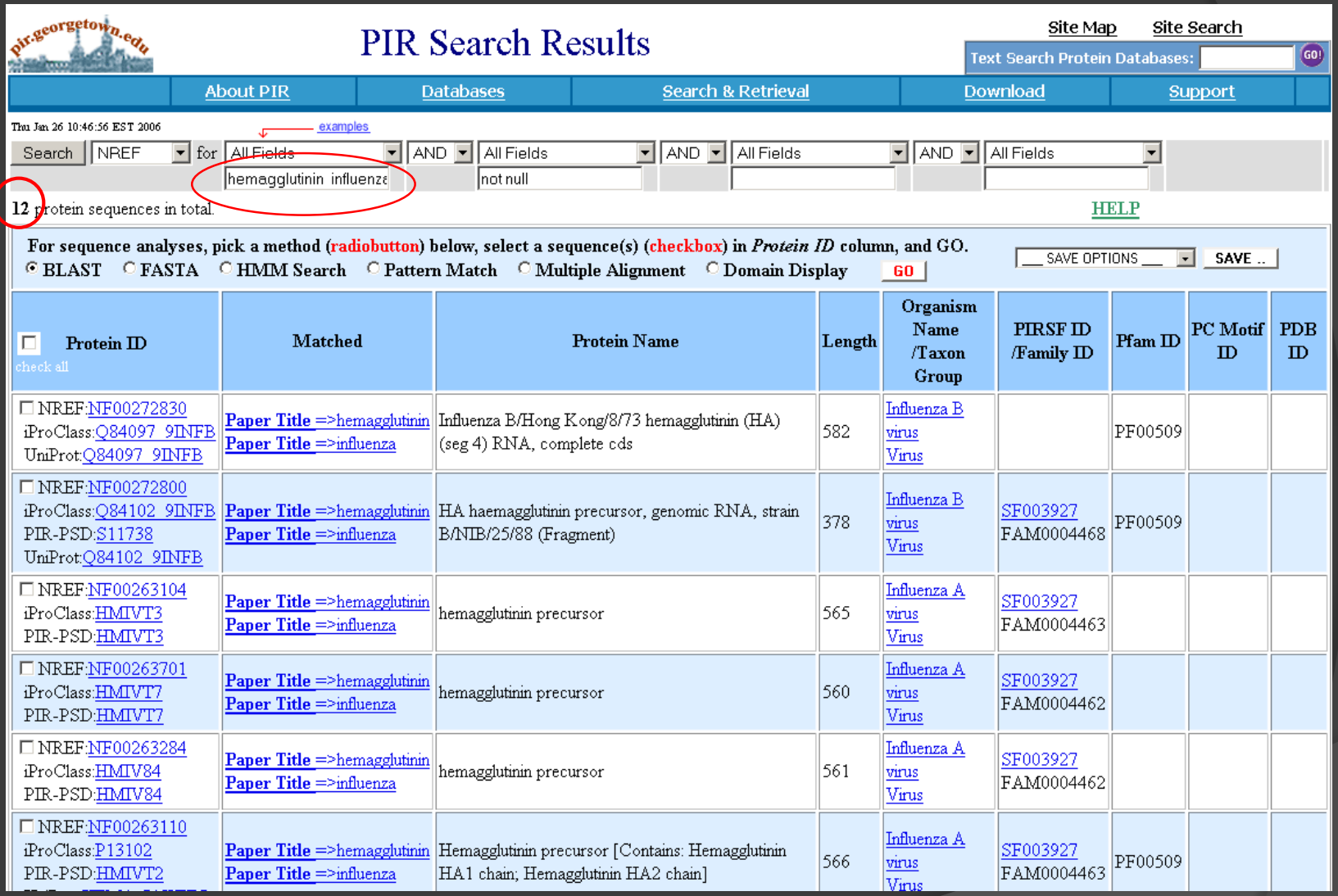
HELP

For sequence analyses, pick a method (radiobutton) below, select a sequence(s) (checkbox) in *Protein ID* column, and GO.

BLAST FASTA HMM Search Pattern Match Multiple Alignment Domain Display GO

SAVE OPTIONS SAVE ..

Protein ID (check all)	Matched	Protein Name	Length	Organism Name /Taxon Group	PIRSF ID /Family ID	Pfam ID	PC Motif ID	PDB ID
NREF:NF02311106 iProClass:Q5G7N6_9CAUD UniProt:Q5G7N6_9CAUD	**Protein Name**=>hemagglutinin **Protein Name** =>hemagglutinin	Putative hemagglutinin protein	801	Listonella pelagia phage phiHSIC Virus				
NREF:NF00058625 iProClass:O55889_9PARA UniProt:O55889_9PARA	**Paper Title** =>hemagglutinin-neuraminidase	RNA dependent RNA polymerase (Fragment)	916	Recombinant PIV3/PIV1 virus Virus	SF000830	PF00946		
NREF:NF00058624 iProClass:Q77Z22_9PARA UniProt:Q77Z22_9PARA	**Protein Name**=>Hemagglutinin **Keyword**=>hemagglutinin **PIRSF Name**=>hemagglutinin-neuraminidase **Pfam Name**=>Hemagglutinin-neuraminidase **Paper Title** =>hemagglutinin-neuraminidase	Hemagglutinin	575	Recombinant PIV3/PIV1 virus Virus	SF001072	PF00423		
NREF:NF00058623 iProClass:O55888_9PARA UniProt:O55888_9PARA	**Paper Title** =>hemagglutinin-neuraminidase	Fusion glycoprotein	555	Recombinant PIV3/PIV1 virus Virus		PF00523		
NREF:NF00244898 iProClass:Q83350_9PARA UniProt:Q83350_9PARA	**Keyword**=>hemagglutinin **PIRSF Name**=>hemagglutinin **Pfam Name**=>Hemagglutinin-neuraminidase	Haemagglutinin protein (H)	607	Morbillivirus Virus	SF003926	PF00423		
NREF:NF00244895 iProClass:Q83352_9PARA UniProt:Q83352_9PARA	**Keyword**=>hemagglutinin **PIRSF Name**=>hemagglutinin **Pfam Name**=>Hemagglutinin-neuraminidase	Haemagglutinin protein (H)	607	Morbillivirus Virus	SF003926	PF00423		



pir.georgetown.edu

PIR Search Results

Site Map Site Search

Text Search Protein Databases: GO

About PIR | Databases | Search & Retrieval | Download | Support

Thu Jan 26 10:46:56 EST 2006

examples

Search NREF for All Fields AND All Fields AND All Fields AND All Fields

hemagglutinin influenza | not null

12 protein sequences in total.

HELP

For sequence analyses, pick a method (radiobutton) below, select a sequence(s) (checkbox) in *Protein ID* column, and GO.

SAVE OPTIONS SAVE ..

BLAST FASTA HMM Search Pattern Match Multiple Alignment Domain Display GO

Protein ID (check all)	Matched	Protein Name	Length	Organism Name /Taxon Group	PIRSF ID /Family ID	Pfam ID	PC Motif ID	PDB ID
NREF:NF00272830 iProClass:Q84097_9INFB UniProt:Q84097_9INFB	Paper Title =>hemagglutinin Paper Title =>influenza	Influenza B/Hong Kong/8/73 hemagglutinin (HA) (seg 4) RNA, complete cds	582	Influenza B virus Virus		PF00509		
NREF:NF00272800 iProClass:Q84102_9INFB PIR-PSD:S11738 UniProt:Q84102_9INFB	Paper Title =>hemagglutinin Paper Title =>influenza	HA haemagglutinin precursor, genomic RNA, strain B/NIB/25/88 (Fragment)	378	Influenza B virus Virus	SF003927 FAM0004468	PF00509		
NREF:NF00263104 iProClass:HMIVT3 PIR-PSD:HMIVT3	Paper Title =>hemagglutinin Paper Title =>influenza	hemagglutinin precursor	565	Influenza A virus Virus	SF003927 FAM0004463			
NREF:NF00263701 iProClass:HMIVT7 PIR-PSD:HMIVT7	Paper Title =>hemagglutinin Paper Title =>influenza	hemagglutinin precursor	560	Influenza A virus Virus	SF003927 FAM0004462			
NREF:NF00263284 iProClass:HMIV84 PIR-PSD:HMIV84	Paper Title =>hemagglutinin Paper Title =>influenza	hemagglutinin precursor	561	Influenza A virus Virus	SF003927 FAM0004462			
NREF:NF00263110 iProClass:P13102 PIR-PSD:HMIVT2	Paper Title =>hemagglutinin Paper Title =>influenza	Hemagglutinin precursor [Contains: Hemagglutinin HA1 chain; Hemagglutinin HA2 chain]	566	Influenza A virus Virus	SF003927 FAM0004463	PF00509		

PIR BLAST/SSearch Results

pir.georgetown.edu

Site Map | Site Search

Text Search Protein Databases: GO!

About PIR | Databases | Search & Retrieval | Download | Support

Sort by E-value | Re-Load | View Superfamily Summary | Search Protein Name

Query sequence:

```
      >Your input sequence:
  1   QDLPGNDNSTATLCLGHHAVPNGTLVKTITDDQIEVTNATELVQSSSTGKICNNPHRILDGIDCTLIDALLGDPHCDVFQ
 81   NETWDLFVERSKAFSNCYPYDVPDYASLRSLVASSGTLEFITEGFTWTGVTQNGRSNACKRGPGSGFFSRLNWLTKSGST
161   YPVLNVTMPNNDNFDKLYIWGIHHPSTNQEQTSLYVQASGRVTVSTRRSQQTIIPNIGSRPWVRGLSSRISIYWTIVKPG
241   DVLVINSNGNLIAPRGYFKMRTGKSSIMRSDAPIDTCISECITPNGSIPNDKPFQNVNKITYGACPKYVKQNTLKLATGM
321   RNVPEKQTGLFGAIAGFIENGWEGMIDGWYGFRHQNSEGTGQAADLKSTQAAIDQINGKLNRVIEKTNEKFHQIEKEFSE
401   VEGRIQDLEKYVEDTKIDLWSYNAELLVALENQHTIDLTDSEMNKLFEKTRRQLRENAEEMGNGCFKIYHKCDNACIESI
481   RNGTYDHDVYRDEALNNRFQIKG
```

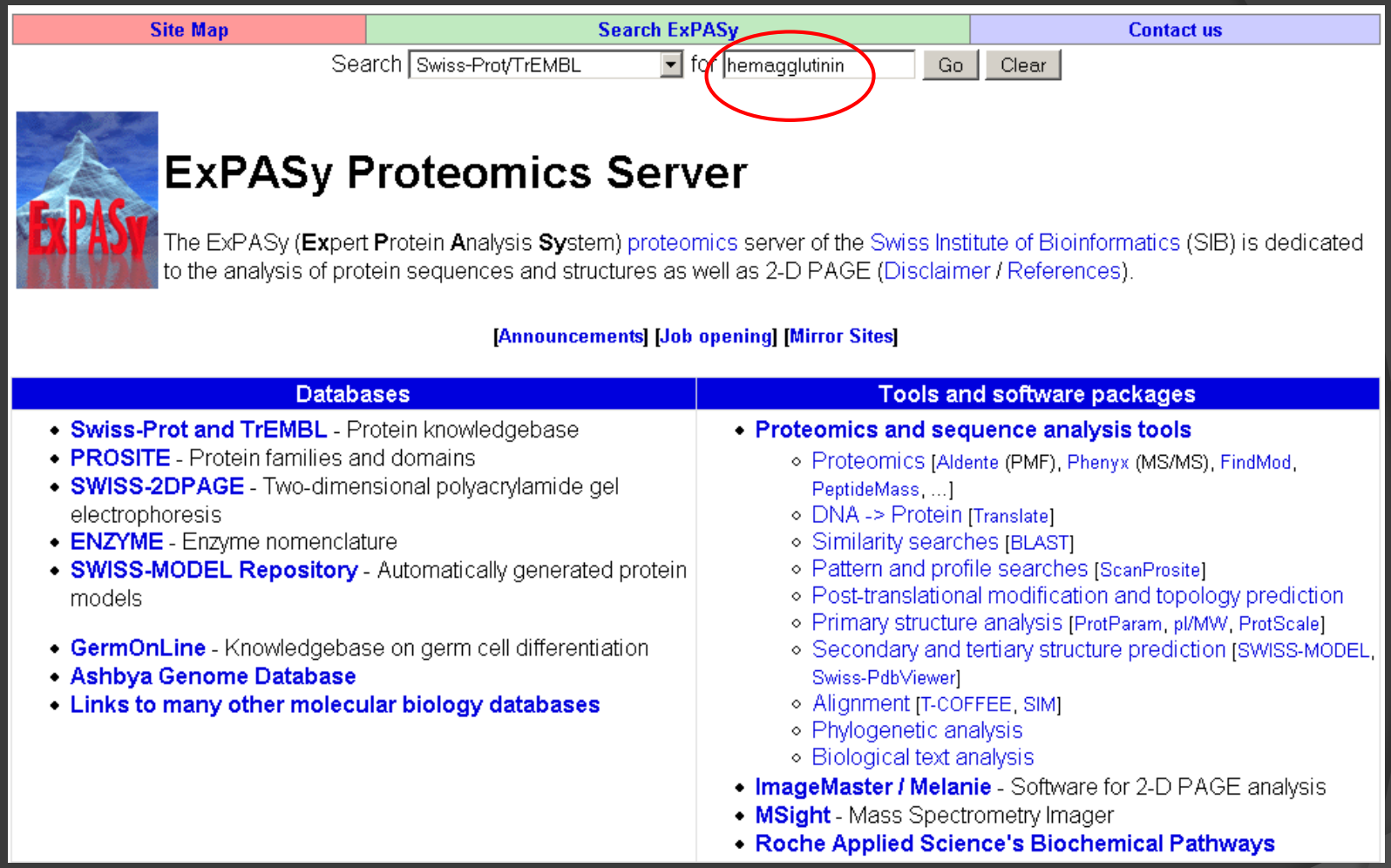
200 match(es) shown in the following table: HELP

For sequence analyses, pick a method (radiobutton) below, select a sequence(s) (checkbox) in *Protein ID* column, and GO.

◉ BLAST ○ FASTA ○ HMM Search ○ Pattern Match ○ Multiple Alignment ○ Domain Display GO

☐ check all Protein ID	Protein Name	Organism	Taxon Group	PIRSF ID	e-value	Length	Ov.lap	%idn	Query Sequence
☐ NREF:NF00263909 iProClass:Q3AVJ6_9SYNE PIR-PSD:HMIVHA UniProt:HEMA_IAAIC	Hemagglutinin precursor [Contains: Hemagglutinin HA1 chain; Hemagglutinin HA2 chain]	Influenza A virus	Virus	SF003927	0.0	566	504	99	
☐ NREF:NF00262744 iProClass:Q5K598_SALSA UniProt:Q67132_9INFA	Hemagglutinin	Influenza A virus	Virus		0.0	566	504	99	
☐ NREF:NF01662375 iProClass:Q47Z20_COLP3 UniProt:Q91MA7_IAHO1	Hemagglutinin	Influenza A virus	Virus		0.0	566	504	99	
☐ NREF:NF01661952 iProClass:Q487E6_COLP3 UniProt:Q910M5_IAHO1	Hemagglutinin	Influenza A virus	Virus		0.0	566	504	98	
☐ NREF:NF00263705 iProClass:Q4ZE13_9CAUD	Hemagglutinin precursor [Contains: Hemagglutinin HA1 chain;	Influenza A	Virus	SF003927	0.0	566	504	98	

```
>NF00263909 Hemagglutinin precursor [Contains: Hemagglutinin HA1
          chain; Hemagglutinin HA2 chain] [Influenza A virus]
         Length = 566

 Score = 1034 bits (2673), Expect = 0.0
 Identities = 502/504 (99%), Positives = 502/504 (99%), Gaps = 1/504 (0%)

Query: 1   QDLPGNDNSTATLCLGHHAVPNGTLVKTITDDQIEVTNATELVQSSSTGKICNNPHRILD 60
           QDLPGNDNSTATLCLGHHAVPNGTLVKTITDDQIEVTNATELVQSSSTGKICNNPHRILD
Sbjct: 17  QDLPGNDNSTATLCLGHHAVPNGTLVKTITDDQIEVTNATELVQSSSTGKICNNPHRILD 76

Query: 61  GIDCTLIDALLGDPHCDVFQNETWDLFVERSKAFSNCYPYDVPDYASLRSLVASSGTLEF 120
           GIDCTLIDALLGDPHCDVFQNETWDLFVERSKAFSNCYPYDVPDYASLRSLVASSGTLEF
Sbjct: 77  GIDCTLIDALLGDPHCDVFQNETWDLFVERSKAFSNCYPYDVPDYASLRSLVASSGTLEF 136

Query: 121 ITEGFTWTGVTQNGRSNACKRGPGSGFFSRLNWLTKSGSTYPVLNVTMPNNDNFDKLYIW 180
           ITEGFTWTGVTQNG SNACKRGPGSGFFSRLNWLTKSGSTYPVLNVTMPNNDNFDKLYIW
Sbjct: 137 ITEGFTWTGVTQNGGSNACKRGPGSGFFSRLNWLTKSGSTYPVLNVTMPNNDNFDKLYIW 196

Query: 181 GIHHPSTNQEQTSLYVQASGRVTVSTRRSQQTIIPNIGSRPWVRGLSSRISIYWTIVKPG 240
           GIHHPSTNQEQTSLYVQASGRVTVSTRRSQQTIIPNIGSRPWVRGLSSRISIYWTIVKPG
Sbjct: 197 GIHHPSTNQEQTSLYVQASGRVTVSTRRSQQTIIPNIGSRPWVRGLSSRISIYWTIVKPG 256

Query: 241 DVLVINSNGNLIAPRGYFKMRTGKSSIMRSDAPIDTCISECITPNGSIPNDKPFQNVNKI 300
           DVLVINSNGNLIAPRGYFKMRTGKSSIMRSDAPIDTCISECITPNGSIPNDKPFQNVNKI
Sbjct: 257 DVLVINSNGNLIAPRGYFKMRTGKSSIMRSDAPIDTCISECITPNGSIPNDKPFQNVNKI 316

Query: 301 TYGACPKYVKQNTLKLATGMRNVPEKQT-GLFGAIAGFIENGWEGMIDGWYGFRHQNSEG 359
           TYGACPKYVKQNTLKLATGMRNVPEKQT GLFGAIAGFIENGWEGMIDGWYGFRHQNSEG
Sbjct: 317 TYGACPKYVKQNTLKLATGMRNVPEKQTRGLFGAIAGFIENGWEGMIDGWYGFRHQNSEG 376

Query: 360 TGQAADLKSTQAAIDQINGKLNRVIEKTNEKFHQIEKEFSEVEGRIQDLEKYVEDTKIDL 419
           TGQAADLKSTQAAIDQINGKLNRVIEKTNEKFHQIEKEFSEVEGRIQDLEKYVEDTKIDL
Sbjct: 377 TGQAADLKSTQAAIDQINGKLNRVIEKTNEKFHQIEKEFSEVEGRIQDLEKYVEDTKIDL 436

Query: 420 WSYNAELLVALENQHTIDLTDSEMNKLFEKTRRQLRENAEEMGNGCFKIYHKCDNACIES 479
           WSYNAELLVALENQHTIDLTDSEMNKLFEKTRRQLRENAEEMGNGCFKIYHKCDNACIES
Sbjct: 437 WSYNAELLVALENQHTIDLTDSEMNKLFEKTRRQLRENAEEMGNGCFKIYHKCDNACIES 496

Query: 480 IRNGTYDHDVYRDEALNNRFQIKG 503
           IRNGTYDHDVYRDEALNNRFQIKG
Sbjct: 497 IRNGTYDHDVYRDEALNNRFQIKG 520
```



Sistema di ricerca e analisi delle informazioni dello Swiss Institute of Bioinformatics in collaborazione con EBI (European Bioinformatics Institute gestendo anche le banche dati SwissProt e TREMBL traduzione delle sequenze nucleotidiche di EMBL

Search in UniProt Knowledgebase (Swiss-Prot and TrEMBL) for: hemagglutinin

UniProtKB/Swiss-Prot Release 48.9 of 24-Jan-2006
UniProtKB/TrEMBL Release 31.9 of 24-Jan-2006

- Number of sequences found in UniProt Knowledgebase (Swiss-Prot(326) and TrEMBL(6933)): **7259**
- Note that the selected sequences can be saved to a file to be later retrieved; to do so, go to the bottom of this page.
- For more directed searches, you can use the Sequence Retrieval System SRS.

Search in UniProtKB/Swiss-Prot: There are matches to 326 out of 206586 entries

AFAF_ECOLI (**Q47037**)
Dr hemagglutinin AFA-III operon regulatory protein afaF. {GENE: Name=afaF} - Escherichia coli
FHAB_BORPE (**P12255**)
Filamentous hemagglutinin. {GENE: Name=fhaB; OrderedLocusNames=BP1879} - Bordetella pertussis
FMA1_ECOLI (**P08180**)
Afimbrial adhesin AFA-I precursor (Dr hemagglutinin AFA-I). {GENE: Name=afaE1; Synonyms=afaE, afaE-1} - Escherichia coli
FMDR_ECOLI (**P24093**)
Dr hemagglutinin structural subunit precursor. {GENE: Name=draA} - Escherichia coli
HA17_CLOBO (**P46083**)
Hemagglutinin component HA-17 (HA 17 kDa subunit). {GENE: Name=HA-17; Synonyms=antP-17} - Clostridium botulinum
HA33_CLOBO (**P46084**)
Main hemagglutinin component (HA 33 kDa subunit) (HA1). {GENE: Name=HA-33; Synonyms=antP-33, ha1} - Clostridium botulinum

ExPASy Home page | Site Map | Search ExPASy | Contact us | Swiss-Prot

Hosted by SIB Switzerland | Mirror sites: Australia | Brazil | Canada | Korea | Taiwan | USA

Search Swiss-Prot/TrEMBL for hemagglutinin influenz Go Clear

Search in UniProt Knowledgebase (Swiss-Prot and TrEMBL) for: hemagglutinin influenza

UniProtKB/Swiss-Prot Release 48.9 of 24-Jan-2006
UniProtKB/TrEMBL Release 31.9 of 24-Jan-2006

- Number of sequences found in UniProt Knowledgebase (Swiss-Prot(177) and TrEMBL)(5408): **5583**
- Note that the selected sequences can be saved to a file to be later retrieved; to do so, go to the bottom of this page.
- For more directed searches, you can use the Sequence Retrieval System SRS.

Search in UniProtKB/Swiss-Prot: There are matches to 177 out of 206586 entries

HEMA_IAAIC (**P03437**)
Hemagglutinin precursor [Contains: Hemagglutinin HA1 chain; Hemagglutinin HA2 chain]. {GENE: Name=HA} - Influenza A virus (strain A/Aichi/2/68 H3N2)

HEMA_IABAN (**P03441**)
Hemagglutinin precursor [Contains: Hemagglutinin HA1 chain; Hemagglutinin HA2 chain] (Fragment). {GENE: Name=HA} - Influenza A virus (strain A/Bangkok/1/79 H3N2)

HEMA_IABUD (**P19694**)
Hemagglutinin precursor [Contains: Hemagglutinin HA1 chain; Hemagglutinin HA2 chain]. {GENE: Name=HA} - Influenza A virus (strain A/Budgerigar/Hokkaido/1/77 H4N6)

HEMA_IACAO (**P26142**)
Hemagglutinin (Fragment). {GENE: Name=HA} - Influenza A virus (strain A/Camel/Mongolia/82 H1N1)

HEMA_IACKA (**P19695**)
Hemagglutinin precursor [Contains: Hemagglutinin HA1 chain; Hemagglutinin HA2 chain]. {GENE: Name=HA} - Influenza A virus (strain A/Chicken/Alabama/1/75)

ExPASy Home page | Site Map | Search ExPASy | Contact us | Swiss-Prot

Search Swiss-Prot/TrEMBL for hemagglutinin influenz Go Clear

UniProtKB/Swiss-Prot entry P03437

Printer-friendly view | Submit update | Quick BlastP search

[Entry info] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

Entry information	
Entry name	**HEMA_IAAIC**
Primary accession number	**P03437**
Secondary accession numbers	None
Entered in Swiss-Prot in	Release 01, July 1986
Sequence was last modified in	Release 01, July 1986
Annotations were last modified in	Release 49, February 2006
Name and origin of the protein	

Cross-references	
Sequence databases	
EMBL	J02090; AAA43178.1; -; Genomic_RNA. [EMBL / GenBank / DDBJ] [CoDingSequence] V01085; CAA24269.1; -; Genomic_RNA.[EMBL / GenBank / DDBJ] [CoDingSequence]
PIR	A93231; HMIVHA.
3D structure databases	
PDB	1EO8; X-ray; A=17-344, B=346-520. [ExPASy / RCSB / EBI] 1HA0; X-ray; A=25-518. [ExPASy / RCSB / EBI] 1HGD; X-ray; A/C/E=17-344, B/D/F=346-520.[ExPASy / RCSB / EBI] 1HGE; X-ray; A/C/E=17-344, B/D/F=346-520.[ExPASy / RCSB / EBI] 1HGF; X-ray; A/C/E=17-344, B/D/F=346-520.[ExPASy / RCSB / EBI] 1HGG; X-ray; A/C/E=17-344, B/D/F=346-520.[ExPASy / RCSB / EBI] 1HGH; X-ray; A/C/E=17-344, B/D/F=346-520.[ExPASy / RCSB / EBI] 1HGI; X-ray; A/C/E=17-344, B/D/F=346-520. [ExPASy / RCSB / EBI] 1HGJ; X-ray; A/C/E=17-344, B/D/F=346-520. [ExPASy / RCSB / EBI] 1HTM; X-ray; A/C/E=17-43, B/D/F=383-520. [ExPASy / RCSB / EBI] 1J8H; X-ray; C=322-334. [ExPASy / RCSB / EBI] 1KEN; X-ray; A/C/E=17-344, B/D/F=346-520.[ExPASy / RCSB / EBI] 1QFU; X-ray; A=17-344, B=346-520. [ExPASy / RCSB / EBI] 1QU1; X-ray; A/B/C/D/E/F=376-530. [ExPASy / RCSB / EBI] 2HMG; X-ray; A/C/E=17-344, B/D/F=346-520.[ExPASy / RCSB / EBI] 2VIR; X-ray; C=44-325. [ExPASy / RCSB / EBI]
Family and domain databases	
InterPro	IPR001364; Hemagglutn. IPR000149; Hemagglutn_1. Graphical view of domain structure.
Pfam	PF00509; Hemagglutinin; 1. Pfam graphical view of domain structure.
PRINTS	PR00330; HEMAGGLUTN1. PR00329; HEMAGGLUTN12.
ProDom	PD000225; Hemagglutn; 1. [Domain structure / List of seq. sharing at least 1 domain]
BLOCKS	P03437.
Other	

Features

 Feature table viewer

 Feature aligner

Key	From	To	Length	Description	FTId
SIGNAL	1	16	16		
CHAIN	17	344	328	Hemagglutinin HA1 chain.	PRO_0000038885
CHAIN	346	566	221	Hemagglutinin HA2 chain.	PRO_0000038886
TOPO_DOM	17	530	514	Extracellular *(Potential)*.	
TRANSMEM	531	551	21	*Potential*.	
TOPO_DOM	552	566	15	Cytoplasmic *(Potential)*.	
LIPID	555	555		S-palmitoyl cysteine (by host) *(By similarity)*.	
LIPID	562	562		S-palmitoyl cysteine (by host) *(By similarity)*.	
LIPID	565	565		S-palmitoyl cysteine (by host) *(By similarity)*.	
CARBOHYD	499	499		N-linked (GlcNAc...).	
DISULFID	30	482		Interchain (between HA1 and HA2 chains).	
DISULFID	68	293			
DISULFID	80	92			
DISULFID	113	155			
DISULFID	297	321			
DISULFID	489	493			
STRAND	27	35	9		
STRAND	37	38	2		
STRAND	40	42	3		
STRAND	45	46	2		
STRAND	48	48	1		
STRAND	50	53	4		
STRAND	55	57	3		
STRAND	59	60	2		
STRAND	65	72	8		
STRAND	74	76	3		
TURN	78	79	2		
HELIX	82	87	6		
TURN	88	88	1		
HELIX	90	95	6		
TURN	96	97	2		
STRAND	99	99	1		
STRAND	101	105	5		

Sequence information

Length: **566 AA** [This is the length of the unprocessed precursor]	Molecular weight: **63416 Da** [This is the MW of the unprocessed precursor]	CRC64: **E395659C23CAFECA** [This is a checksum on the sequence]

```
        10         20         30         40         50         60
MKTIIALSYI FCLALGQDLP GNDNSTATLC LGHHAVPNGT LVKTITDDQI EVTNATELVQ

        70         80         90        100        110        120
SSSTGKICNN PHRILDGIDC TLIDALLGDP HCDVFQNETW DLFVERSKAF SNCYPYDVPD

       130        140        150        160        170        180
YASLRSLVAS SGTLEFITEG FTWTGVTQNG GSNACKRGPG SGFFSRLNWL TKSGSTYPVL

       190        200        210        220        230        240
NVTMPNNDNF DKLYIWGIHH PSTNQEQTSL YVQASGRVTV STRRSQQTII PNIGSRPWVR

       250        260        270        280        290        300
GLSSRISIYW TIVKPGDVLV INSNGNLIAP RGYFKMRTGK SSIMRSDAPI DTCISECITP

       310        320        330        340        350        360
NGSIPNDKPF QNVNKITYGA CPKYVKQNTL KLATGMRNVP EKQTRGLFGA IAGFIENGWE

       370        380        390        400        410        420
GMIDGWYGFR HQNSEGTGQA ADLKSTQAAI DQINGKLNRV IEKTNEKFHQ IEKEFSEVEG
```

Formato FASTA
formato di interscambio
tra software

```
>sp|P03437|HEMA_IAAIC Hemagglutinin precursor [Contains: Hemagglutinin HA1 chain; Hemagglutinin HA2 c
MKTIIALSYIFCLALGQDLPGNDNSTATLCLGHHAVPNGTLVKTITDDQIEVTNATELVQ
SSSTGKICNNPHRILDGIDCTLIDALLGDPHCDVFQNETWDLFVERSKAFSNCYPYDVPD
YASLRSLVASSGTLEFITEGFTWTGVTQNGGSNACKRGPGSGFFSRLNWLTKSGSTYPVL
NVTMPNNDNFDKLYIWGIHHPSTNQEQTSLYVQASGRVTVSTRRSQQTIIPNIGSRPWVR
GLSSRISIYWTIVKPGDVLVINSNGNLIAPRGYFKMRTGKSSIMRSDAPIDTCISECITP
NGSIPNDKPFQNVNKITYGACPKYVKQNTLKLATGMRNVPEKQTRGLFGAIAGFIENGWE
GMIDGWYGFRHQNSEGTGQAADLKSTQAAIDQINGKLNRVIEKTNEKFHQIEKEFSEVEG
RIQDLEKYVEDTKIDLWSYNAELLVALENQHTIDLTDSEMNKLFEKTRRQLRENAEEMGN
GCFKIYHKCDNACIESIRNGTYDHDVYRDEALNNRFQIKGVELKSGYKDWILWISFAISC
FLLCVVLLGFIMWACQRGNIRCNICI
```

RCSB PDB
PROTEIN DATA BANK

A MEMBER OF THE PDB

An Information Portal to Biological Macromolecular Structures

As of Tuesday Jan 24, 2006 there are 34777 Structures | PDB Statistics

Contact Us | Help | Print Page

All | PDB ID or keyword | Web Pages | Author | hemagglutinin | SEARCH | Advanced Search

Home | Search

- Home
- Tutorial About This Site
- Getting Started
- Download Files
- Deposit and Validate
- Structural Genomics
- Dictionaries & File Formats
- Software Tools
- Educational Resources
- General Information
- Acknowledgements
- Frequently Asked Questions
- Known Problems
- Report Bugs/Comments

Welcome to the RCSB PDB

The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

The RCSB is a member of the wwPDB whose mission is to ensure that the PDB archive remains an international resource with uniform data.

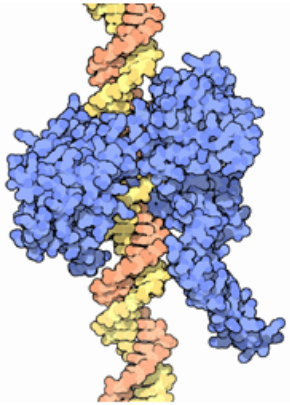
This site offers tools for browsing, searching, and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

Information about compatible browsers can be found here.

A narrated tutorial illustrates how to search, navigate, browse, generate reports and visualize structures using this new site. [This requires the Macromedia Flash player download.]

Comments? info@rcsb.org

Molecule of the Month: Topoisomerases



Each of your cells contains about 2 meters of DNA, all folded into the tiny space inside the nucleus, which is a million times smaller. As you might imagine, these long, thin strands can get tangled very easily in the busy environment of the nucleus. To make things even more complicated, DNA is a double helix, which must be unwound to access the genetic information.

- More ...
- Previous Features

NEWS

- Complete News
- Newsletter
- Discussion Forum

24-Jan-2006

Montville Township High School Places First at the NJ Science Olympiad Protein Modeling Trial Event

Several high school teams competed in the protein modeling event at the New Jersey Northern Regional Science Olympiad that was held January 12, 2006 at Montclair State University.

- Full Story ...

17-Jan-2006

Time-stamped Copies of PDB Archive Available via FTP

10-Jan-2006

Structural Genomics Tools and Portal

RCSB PDB PROTEIN DATA BANK

A MEMBER OF THE PDB

An Information Portal to Biological Macromolecular Structures

As of Tuesday Jan 24, 2006 there are 34777 Structures | PDB Statistics

Contact Us | Help | Print Page

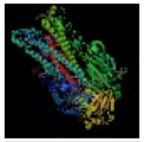
All | PDB ID or keyword | Web Pages | Author | SEARCH | Advanced Search

Home | Search | Results | Queries

- Results (1-10 of 87)
- Refine this Search
- 1 Structures Awaiting Release
- Select All
- Deselect All
- Download Selected
- Tabulate
- Narrow Query
- Sort Results
- Results per Page
- Show Query Details
- Results Help

1 2 3 4 5 .. 9

☑ 1RUZ



1918 H1 Hemagglutinin

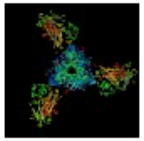
Characteristics **Release Date:** 30-Mar-2004 **Exp. Method:** X Ray Diffraction **Resolution:** 2.90 Å

Classification Virus/viral Protein

Compound **Mol. Id:** 1 **Molecule:** Hemagglutinin **Mol. Id:** 2 **Molecule:** Hemagglutinin

Authors Gamblin, S.J., Haire, L.F., Russell, R.J., Stevens, D.J., Xiao, B., Ha, Y., Vasisht, N., Steinhauer, D.A., Daniels, R.S., Elliot, A., Wiley, D.C., Skehel, J.J.

☑ 1EO8



INFLUENZA VIRUS HEMAGGLUTININ COMPLEXED WITH A NEUTRALIZING ANTIBODY

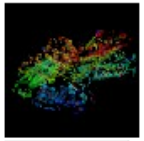
Characteristics **Release Date:** 12-Apr-2000 **Exp. Method:** X Ray Diffraction **Resolution:** 2.80 Å

Classification Virus/viral Protein

Compound **Mol. Id:** 1 **Molecule:** Hemagglutinin (ha1 Chain) **Fragment:** Bromelain Released Fragment **Mol. Id:** 2 **Molecule:** Hemagglutinin (ha2 Chain) **Fragment:** Bromelain Released Fragment **Mol. Id:** 3 **Molecule:** Antibody (light Chain) **Fragment:** Fab Fragment of Antibody Bh151 **Mol. Id:** 4 **Molecule:** Antibody (heavy Chain) **Fragment:** Fab Fragment of Antibody Bh151

Authors Fleury, D., Daniels, R.S., Skehel, J.J., Knossow, M., Bizebard, T.

☑ 1FLC



X-RAY STRUCTURE OF THE HAEMAGGLUTININ-ESTERASE-FUSION GLYCOPROTEIN OF INFLUENZA C VIRUS

Characteristics **Release Date:** 01-Mar-2000 **Exp. Method:** X Ray Diffraction **Resolution:** 3.20 Å

Classification Hydrolase

Compound **Mol. Id:** 1 **Molecule:** Haemagglutinin Esterase Fusion Glycoprotein **Fragment:** Hef1 **Mol. Id:** 2 **Molecule:** Haemagglutinin Esterase Fusion Glycoprotein **Fragment:** Hef2

Authors Rosenthal, P.B., Zhang, X.

☑ 1FYT

CRYSTAL STRUCTURE OF A COMPLEX OF A HUMAN ALPHA/BETA-T CELL RECEPTOR, INFLUENZA HA ANTIGEN PEPTIDE, AND MHC CLASS II MOLECULE, HLA-DR1

Home | Search | Structure | Results | Queries

Structure Summary | Biology & Chemistry | Materials & Methods | Sequence Details | Geometry

- 1RUZ
- Download Files
- FASTA Sequence
- Display Files
- Display Molecule
- Structural Reports
- Structure Analysis
- Help

1RUZ

Title	1918 H1 Hemagglutinin
Authors	Skehel, J.J., Gamblin, S.J., Haire, L.F., Russell, R.J., Stevens, D.J., Xiao, B., Ha, Y., Vasisht, N., Steinhauer, D.A., Daniels, R.S.
Primary Citation	**Gamblin, S.J., Haire, L.F., Russell, R.J., Stevens, D.J., Xiao, B., Ha, Y., Vasisht, N., Steinhauer, D.A., Daniels, R.S., Elliot, A., Wiley, D.C., Skehel, J.J.** The structure and receptor binding properties of the 1918 influenza hemagglutinin. *Science v303 pp.1838-1842 , 2004* [Abstract]
History	Deposition 2003-12-12 Release 2004-03-30
Experimental Method	Type X-RAY DIFFRACTION Data N/A

Parameters

Resolution[Å]	R-Value	R-Free	Space Group
2.90	0.248 (obs.)	0.289	P 4_1 2_1 2

Unit Cell

Length [Å]	a	171.46	b	171.46	c	153.45
Angles [°]	alpha	90.00	beta	90.00	gamma	90.00

Molecular Description

Polymer: 1 Molecule: hemagglutinin Chains: H, J, L; Other Details: Hemagglutinin HA1 chain Polymer: 2 Molecule: hemagglutinin Chains: I, K, M; Other Details: Hemagglutinin HA2 chain

Functional Class

Virus/viral Protein

Source

Polymer: 1 Scientific Name: Influenza a virus Common Name: Virus Polymer: 2 Scientific Name: Influenza a virus Common Name: Virus

Related PDB Entries

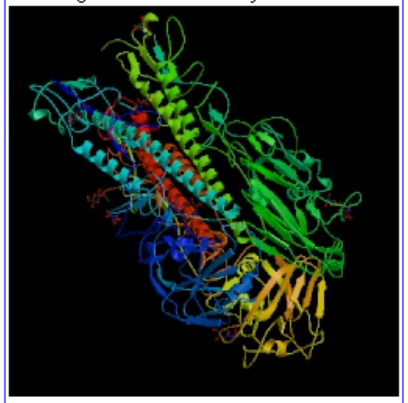
Id	Details
1RVZ	1934 H1 Hemagglutinin in complex with LSTC
1RVX	1934 H1 Hemagglutinin in complex with LSTA
1RVT	1930 H1 Hemagglutinin in complex with LSTC
1RV0	1930 Swine H1 Hemagglutinin complexed with LSTA
1RUY	1930 H1 Hemagglutinin
1RU7	1934 H1 Hemagglutinin

Chemical Component

Identifier	Name	Formula	Drug Similarity	Ligand Structure	Ligand Interaction

Images and Visualization

Biological Molecule / Asymmetric Unit



Display Options

KiNG
Jmol
WebMol
Protein Workshop
QuickPDB
All Images

La qualità e la quantità dei dati hanno spinto gli scienziati a puntare verso traguardi proporzionatamente ambiziosi:

-Comprendere gli aspetti integrativi della biologia degli organismi osservati come sistemi complessi coerenti.

-Correlare sequenza, struttura, interazioni e funzionalità di singole proteine, acidi nucleici e complessi, tra di essi

-Usare dati riguardanti organismi contemporanei come base per viaggiare avanti e indietro nel tempo; per dedurre eventi della storia evoluzionistica.

-Fornire un supporto ad applicazioni nel campo della medicina, agricoltura ecc.

Scenario futuro:

Nuovo Virus altamente pericoloso> pandemie sul pianeta

Sequenziamento del suo genoma > confronto con la sua sequenza con quella dei virus presenti in Banca Dati > Individuazione di proteine patogene per l'uomo responsabili dell'infettività e virulenza > Determiazione della struttura tridimensionale di tali proteine tramite Homology modeling o Modeling predittivo > confronto tramite Banca Dati con struttura di altre proteine e di conseguenza loro funzione > Drug Design per ricerca di nuovi farmaci o individuazioni di Anticorpi che possano neutralizzare il Virus.

Tempo necessario pochi giorni > pandemia evitata

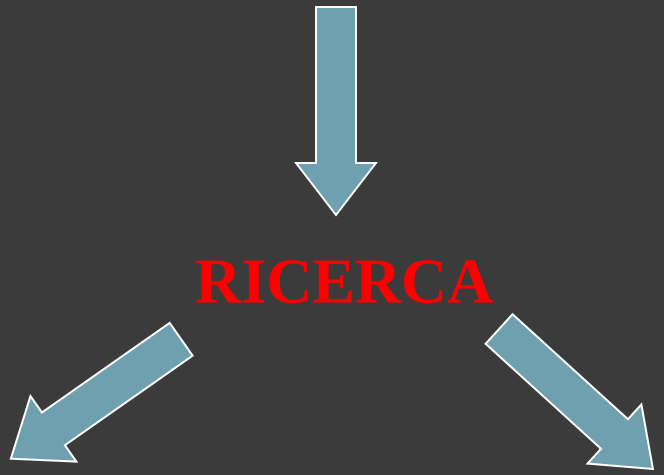
Confronto di sequenze

Il confronto tra sequenze è in biologia computazionale la base per

- misurare la “similarità” tra le sequenze
  - allineamento
- misurare la “diversità” tra le sequenze
  - distanza di edit
- trovare parti comuni alle sequenze
  - pattern discovery
  - allineamento locale

INTRODUZIONE

DATABASE DI SEQUENZE



RICERCA

TESTUALE

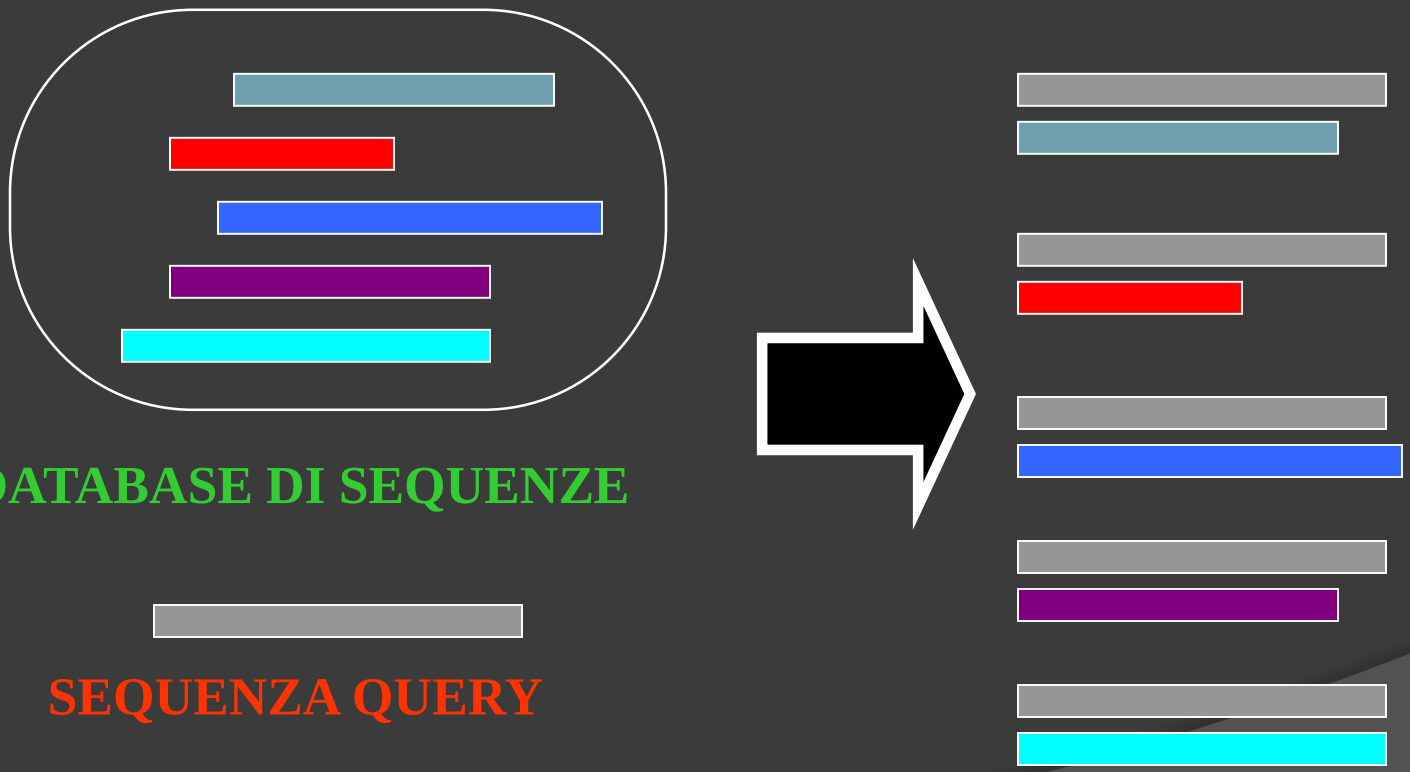
Ricerca dei record i cui campi soddisfano determinati criteri (hanno certi valori)

SIMILARITA'

Ricerca dei record che hanno le sequenze più "simili" ad una sequenza fornita come query

RICERCA PER SIMILARITA’

- La ricerca per similarità di una sequenza contro un database di sequenze richiede che sia possibile valutare la similarità della sequenza query contro ciascuna delle sequenza del database. Quindi il problema da risolvere è quello della ricerca delle similarità tra due sequenze



DATABASE DI SEQUENZE

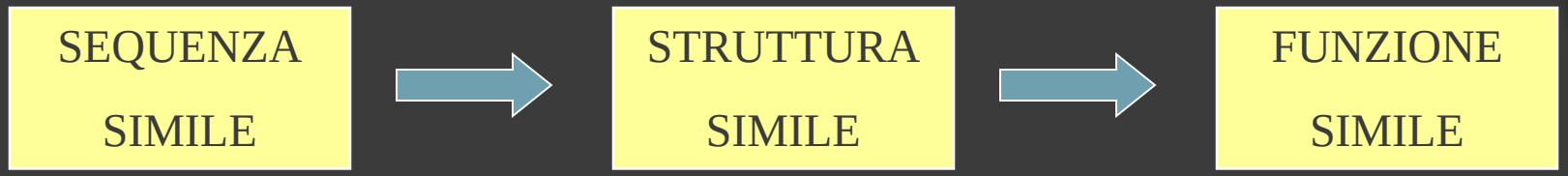
SEQUENZA QUERY

PERCHE' CERCARE SEQUENZE SIMILI?

- Quando si ottiene (in qualche modo) una sequenza di DNA o Aminoacidi si è interessati a capire cos'è quella sequenza (è già nota?) e a scoprire la sua funzione.
- Potrebbe anche capitare che la sequenza stessa sia presente nei database e già annotata (descritta la sua funzione)... Nel caso invece non si trovasse nei database esattamente la stessa sequenza, un modo semplice di ipotizzare (è comunque una predizione, che dovrà poi essere confermata sperimentalmente) la funzione della mia sequenza query è quello di cercare sequenze simili che invece siano già state annotate.
- In base al grado di similarità trovato diventa possibile fare delle ipotesi più o meno probabili sulla funzione della sequenza query semplicemente "trasferendo" ad essa la funzione delle sequenze target simili ad essa identificate .

QUANDO SUPPORRE LA FUNZIONE

- Se le sequenze di due proteine (DNA) sono molto simili allora lo saranno anche le strutture e le funzioni



- Non vale il viceversa! (Funzioni e strutture simili non implicano sequenze simili)

- Ci possono essere proteine con la stessa funzione, ma con struttura e soprattutto sequenza diversa. Es. mutazioni silenti, che interessano la terza base di un codone. L'aminoacido rimane lo stesso ma è cambiato il DNA!

SIMILARITA' E OMOLOGIA

- Spesso si fa confusione tra similarità ed omologia!

- La similarità è un aspetto quantitativo che indica (fissato un criterio comparativo, % identità, % mutazioni conservative...) un livello di somiglianza tra le sequenze.

- L'omologia è un aspetto qualitativo che riguarda più propriamente la "funzione" delle sequenze ed indica un'origine filogenetica comune

Dobbiamo ricorrere a criteri statistici per giudicare la significatività delle similitudini e delle differenze

Proteine omologhe: proteine che si sono evolute da un comune ancestore, nell'evoluzione la similarità di sequenza è meno preservata rispetto alla struttura terziaria

Si possono avere proteine omologhe con un'identità di sequenza fino al 20%

Come è possibile ciò?

La maggior parte delle mutazioni avviene sulla superficie della proteina mentre gli amminoacidi del core sono maggiormente conservati in modo da consentire il medesimo folding alle proteine.

I problemi principali nella deduzione dei rapporti filogenetici mediante confronto tra sequenze sono:

- Ampio ambito di variabilità della similarità, che può scendere al di sotto della significatività statistica
- Effetto delle diverse velocità evolutive lungo i diversi rami dell'albero evolutivo

The Genetic Code

Second letter

First letter	U	C	A	G	
U	UUU, UUC: Phenyl-alanine UUA, UUG: Leucine	UCU, UCC, UCA, UCG: Serine	UAU, UAC: Tyrosine UAA: Stop codon UAG: Stop codon	UGU, UGC: Cysteine UGA: Stop codon UGG: Tryptophan	U C A G
C	CUU, CUC, CUA, CUG: Leucine	CCU, CCC, CCA, CCG: Proline	CAU, CAC: Histidine CAA, CAG: Glutamine	CGU, CGC, CGA, CGG: Arginine	U C A G
A	AUU, AUC, AUA: Isoleucine AUG: Methionine; initiation codon	ACU, ACC, ACA, ACG: Threonine	AAU, AAC: Asparagine AAA, AAG: Lysine	AGU, AGC: Serine AGA, AGG: Arginine	U C A G
G	GUU, GUC, GUA, GUG: Valine	GCU, GCC, GCA, GCG: Alanine	GAU, GAC: Aspartic acid GAA, GAG: Glutamic acid	GGU, GGC, GGA, GGG: Glycine	U C A G

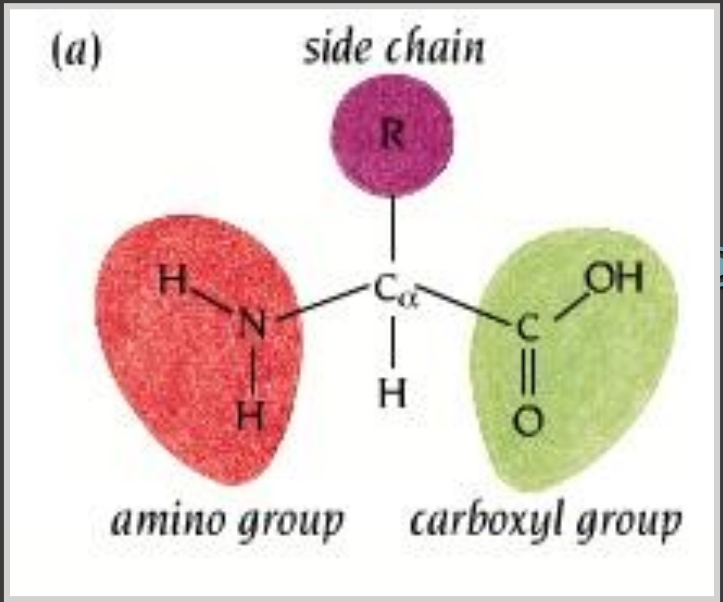
- The genetic code - Each amino acid is coded by 3 nucleotides, named *codon*.
- Code redundancy - Most amino acids are coded by several codons.
  - 64 triplets code for 20 amino acids & 3 stop codons.

PRINCIPI DELLA STRUTTURA DELLE PROTEINE

- Struttura primaria: gli amminoacidi e il legame peptidico
- Struttura secondaria
  - ✓ Strutture supersecondarie o motivi
- Struttura terziaria
  - ✓ Domini
- Struttura quaternaria
- Classificazione del *folding* proteico (SCOP e CATH)
- Il Protein Data Bank (PDB)

STRUTTURA PRIMARIA

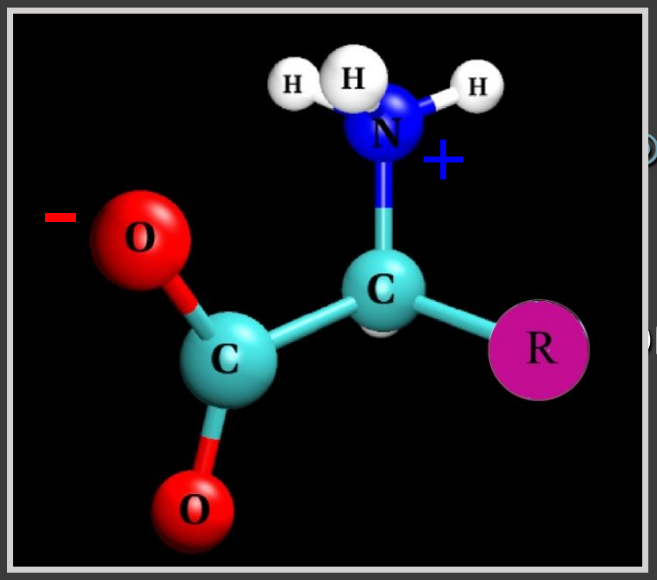
Il codice genetico specifica 20 diverse catene laterali, mentre altre possono essere prodotte dalle prime ad opera di enzimi (modificazioni post-traduzionali)



- 20 α-amminoacidi standard

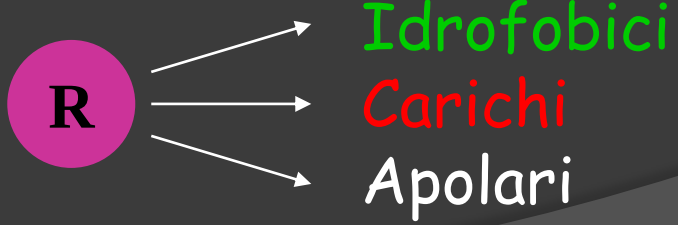
α-amminoacido: gruppo amminico legato al C (α) adiacente al gruppo carbossilico

STRUTTURA PRIMARIA

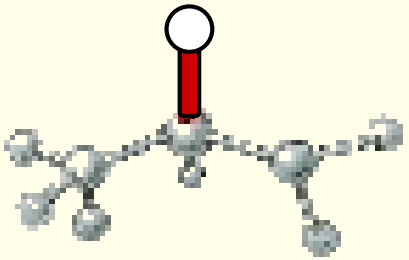


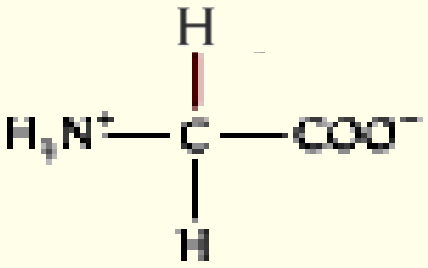
- ph=7 AA anfoteri o Zwitterioni
- roprietà sia basiche che acide

Classificazione: in base alla loro catena laterale



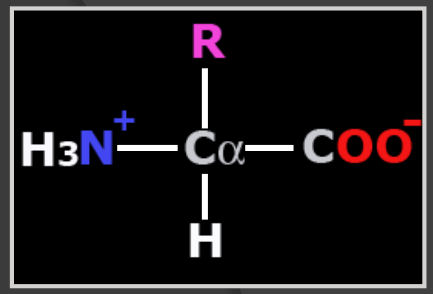
GLICINA

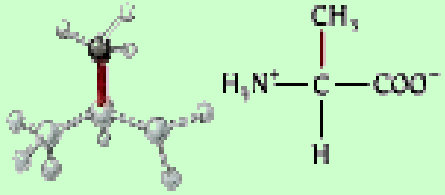




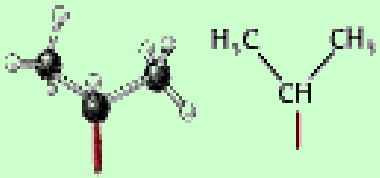
G ***Gly, Glicina***

Amminoacidi IDROFOBICI

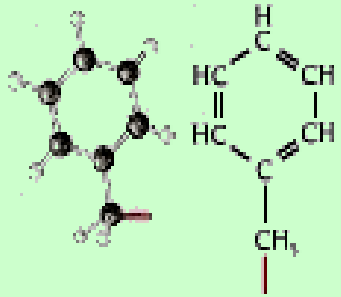




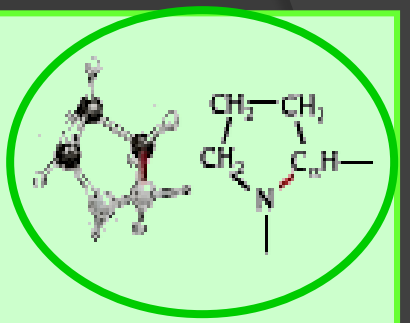
A *Ala, Alanina*



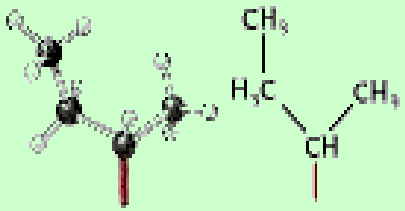
V *Val, Valina*



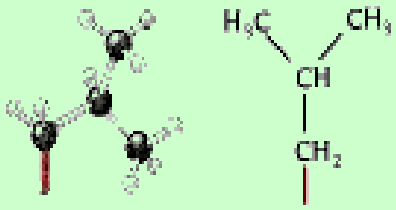
F *Phe, Fenilanina*



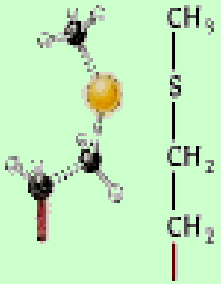
P *Pro, Prolina*



I *Ile, Isoleucina*

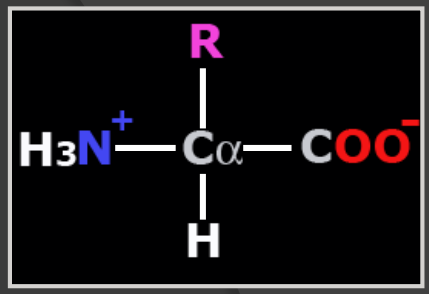


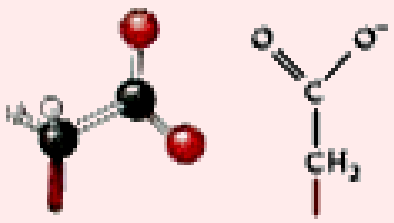
L *Leu, Leucina*



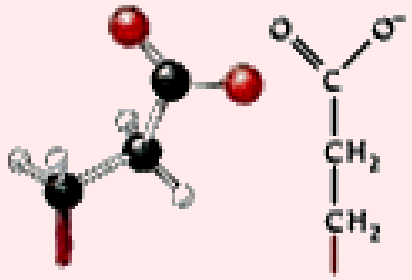
M *Met, Metionina*

Amminoacidi CARICHI

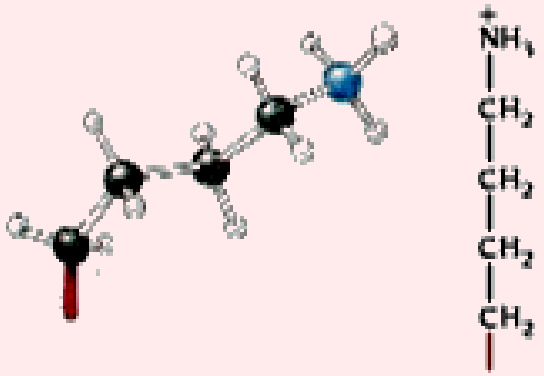




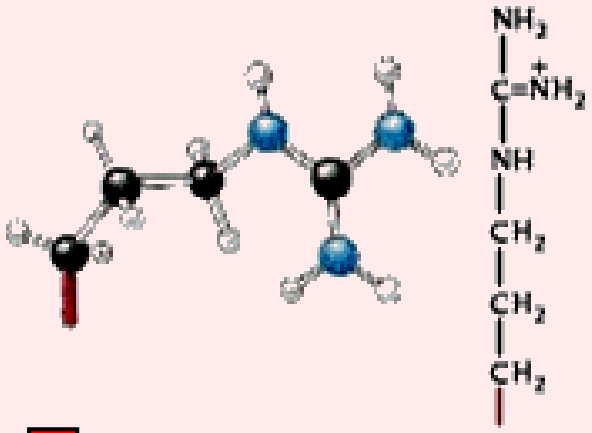
D *Asp, Acido Aspartico* (-)
(Aspartato)



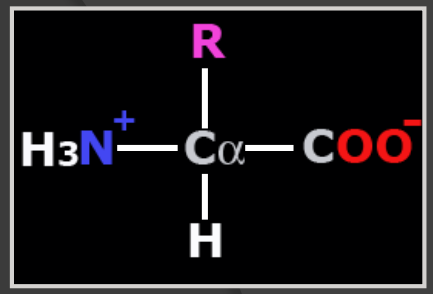
E *Glu, Acido Glutammico* (-)
(Glutammato)



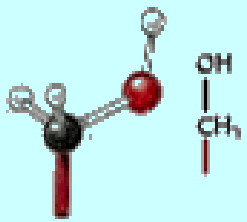
K *Lys, Lisina* (+)



R *Arg, Arginina* (+)

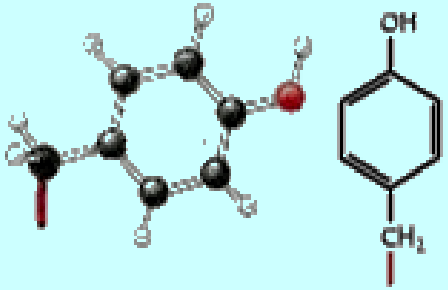


Amminoacidi POLARI

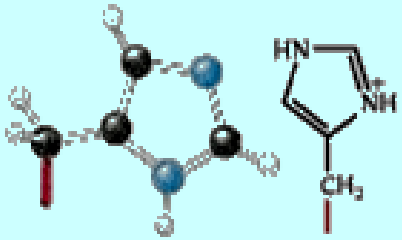


S *Ser, Serina*

T *Thr, Treonina*

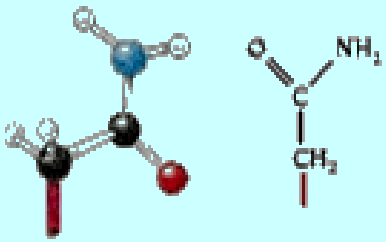


Y *Tyr, Tirosina*

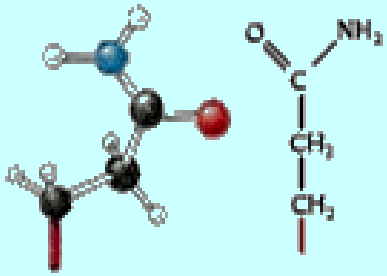


H *His, Istidina*

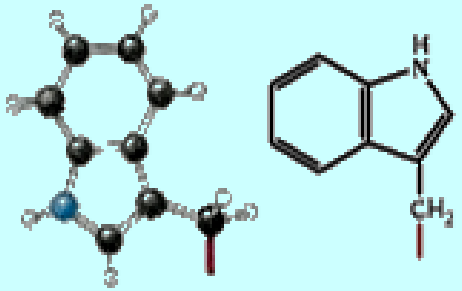
C *Cys, Cisteina*



N *Asn, Asparagina*

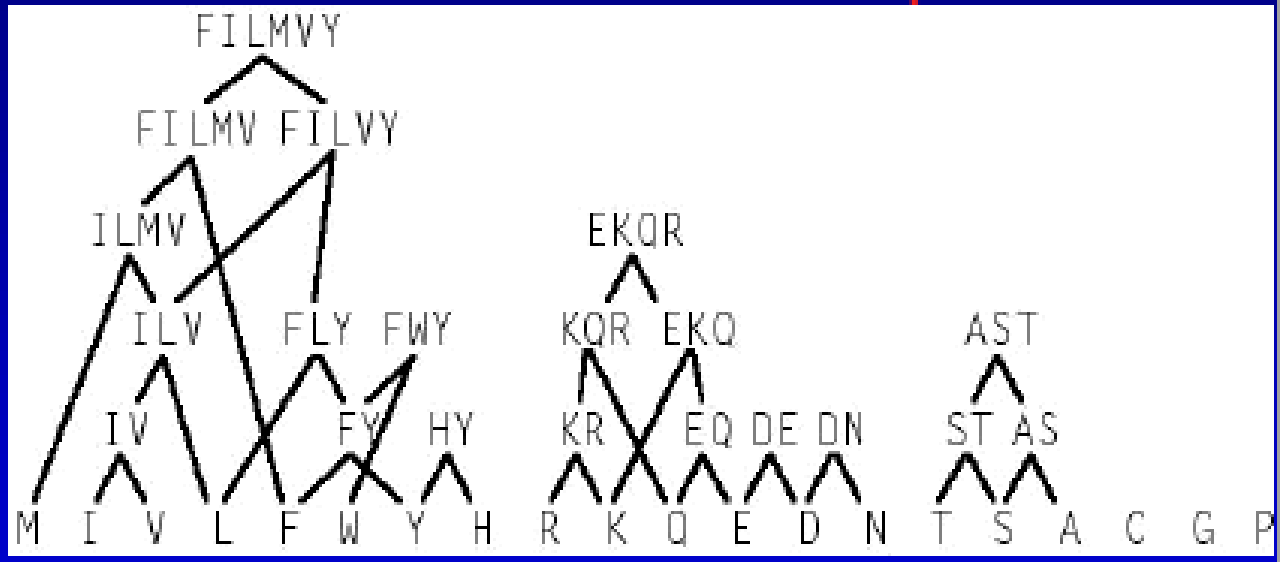


Q *Gln, Glutammina*



W *Trp, Triptofano*

Allowable Amino Acid Substitution Groups



http://www.imb-jena.de/IMAGE_AA.html

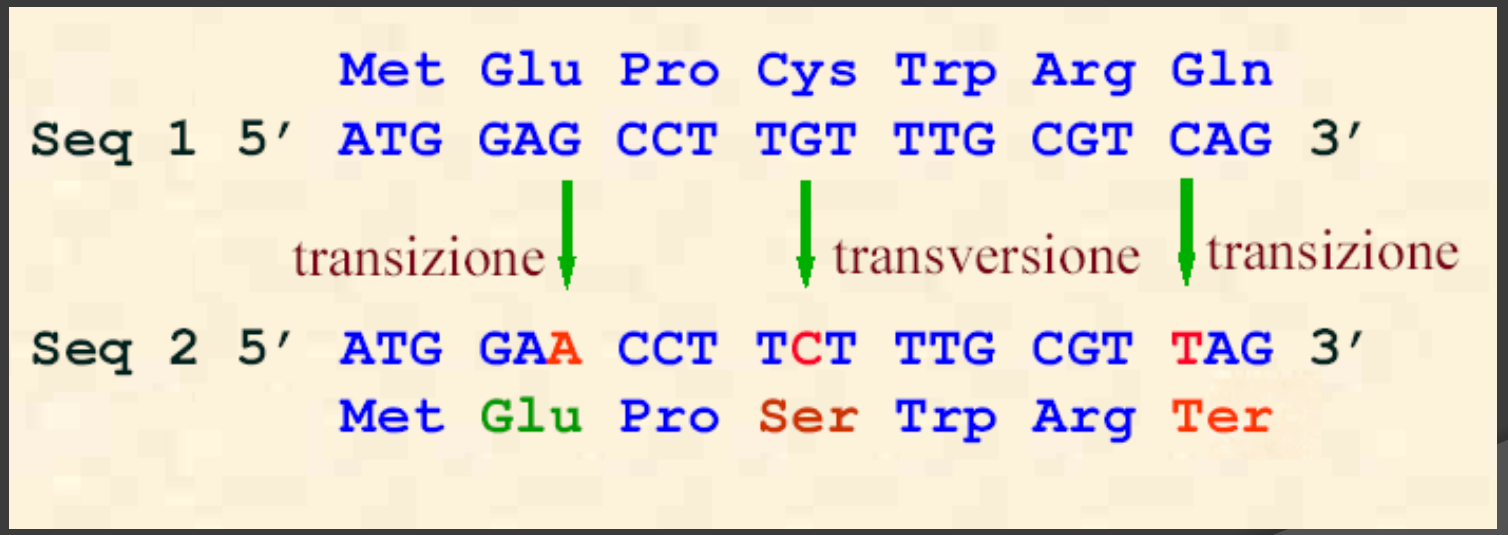
amino acid		
Alanine	ALA	A
Arginine	ARG	R
Aspartic Acid	ASP	D
Asparagine	ASN	N
Cysteine	CYS	C
Glutamic Acid	GLU	E
Glutamine	GLN	Q
Glycine	GLY	G
Histidine	HIS	H
Isoleucine	ILE	I
Leucine	LEU	L
Lysine	LYS	K
Methionine	MET	M
Phenylalanine	PHE	F
Proline	PRO	P
Serine	SER	S
Threonine	THR	T
Tryptophan	TRP	W
Tyrosine	TYR	Y
Valine	VAL	V

EVOLUZIONE DEI GENOMI

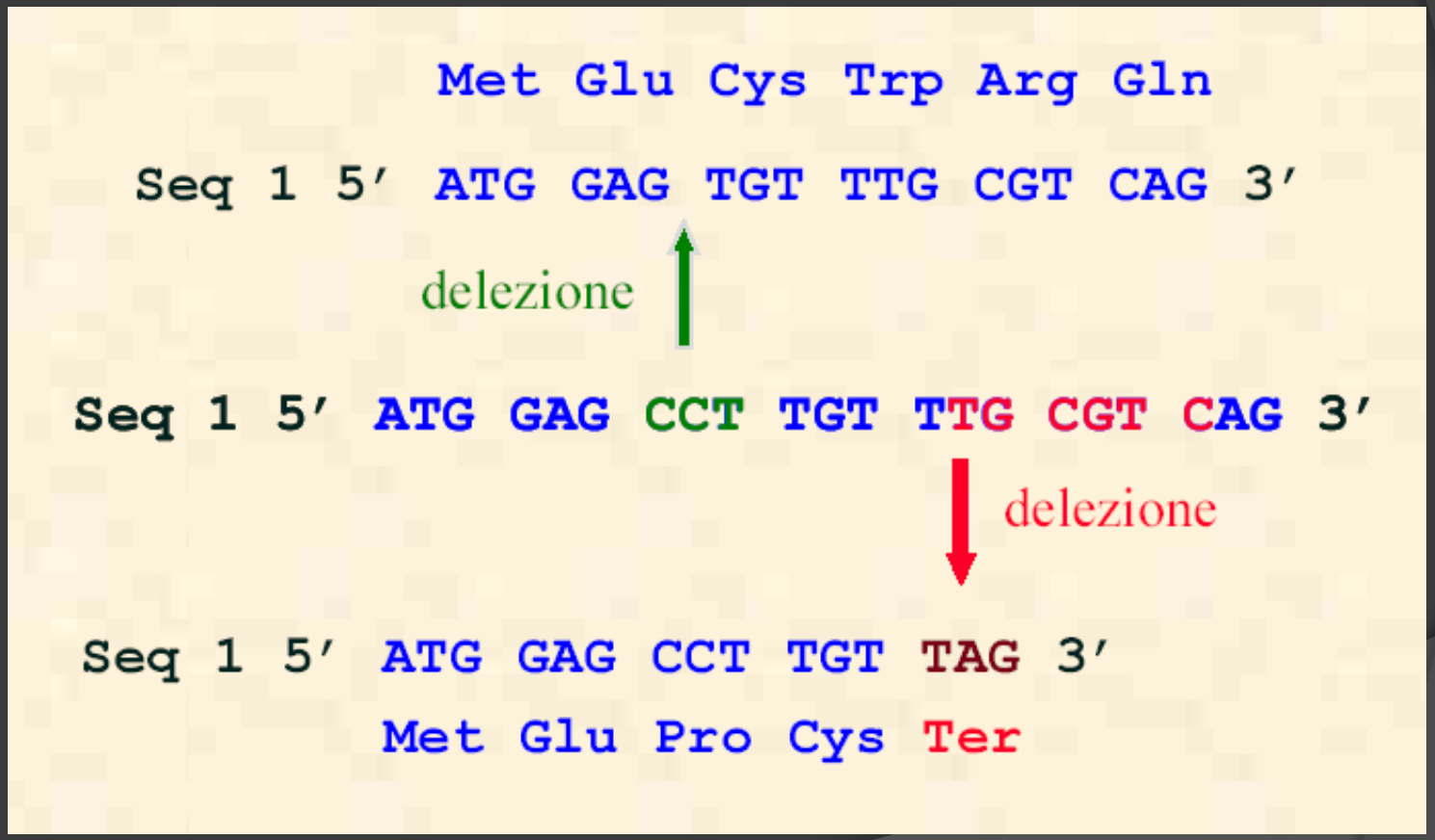
- Vari sono i meccanismi responsabili della variabilità genetica che oggi possiamo osservare:

- Mutazioni puntiformi
- Delezioni
- Inserzioni
- Inversioni

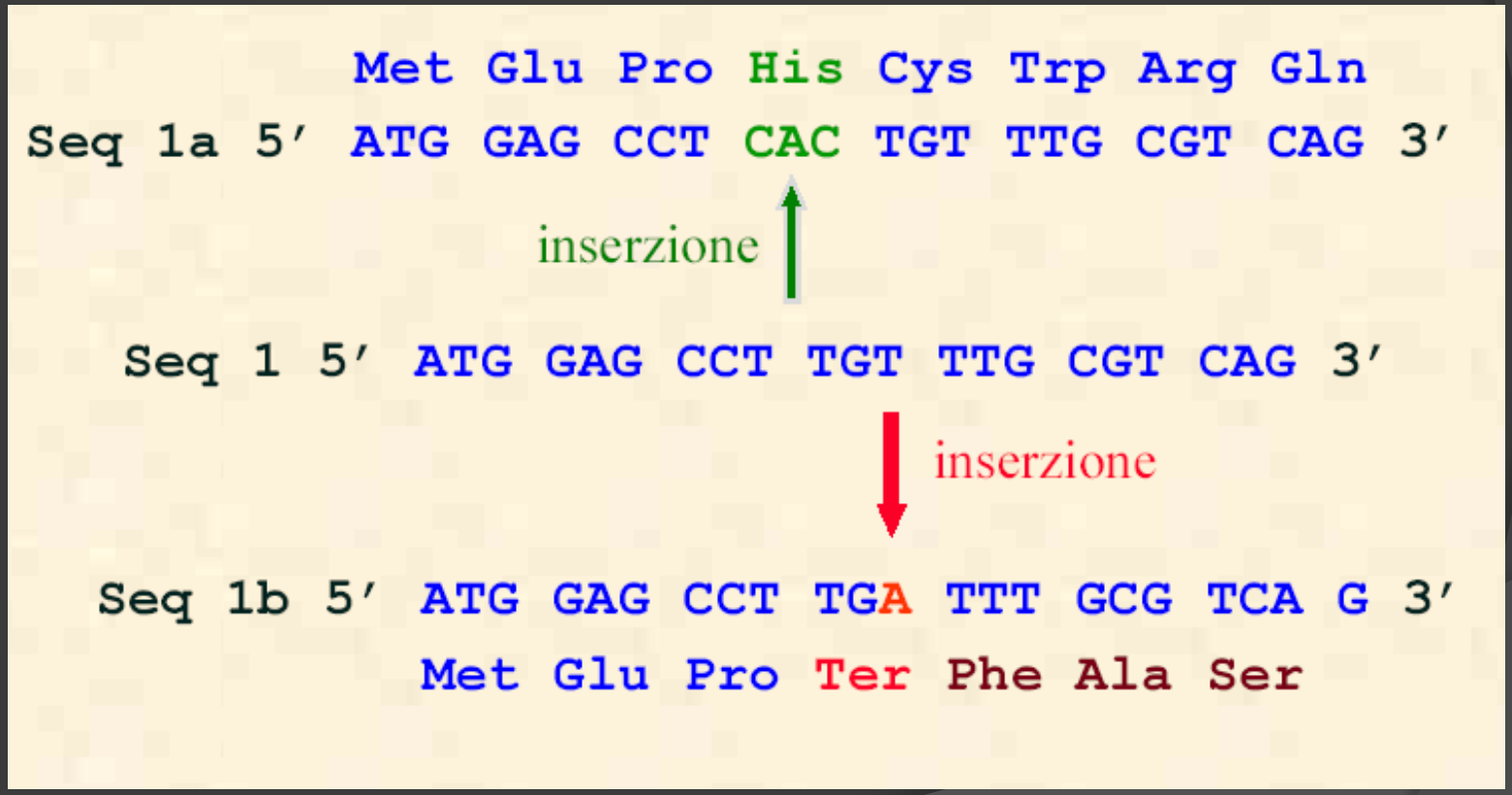
MUTAZIONI


Met Glu Pro Cys Trp Arg Gln
Seq 1 5′ ATG GAG CCT TGT TTG CGT CAG 3′
transizione
transversione
transizione
Seq 2 5′ ATG GAA CCT TCT TTG CGT TAG 3′
Met Glu Pro Ser Trp Arg Ter

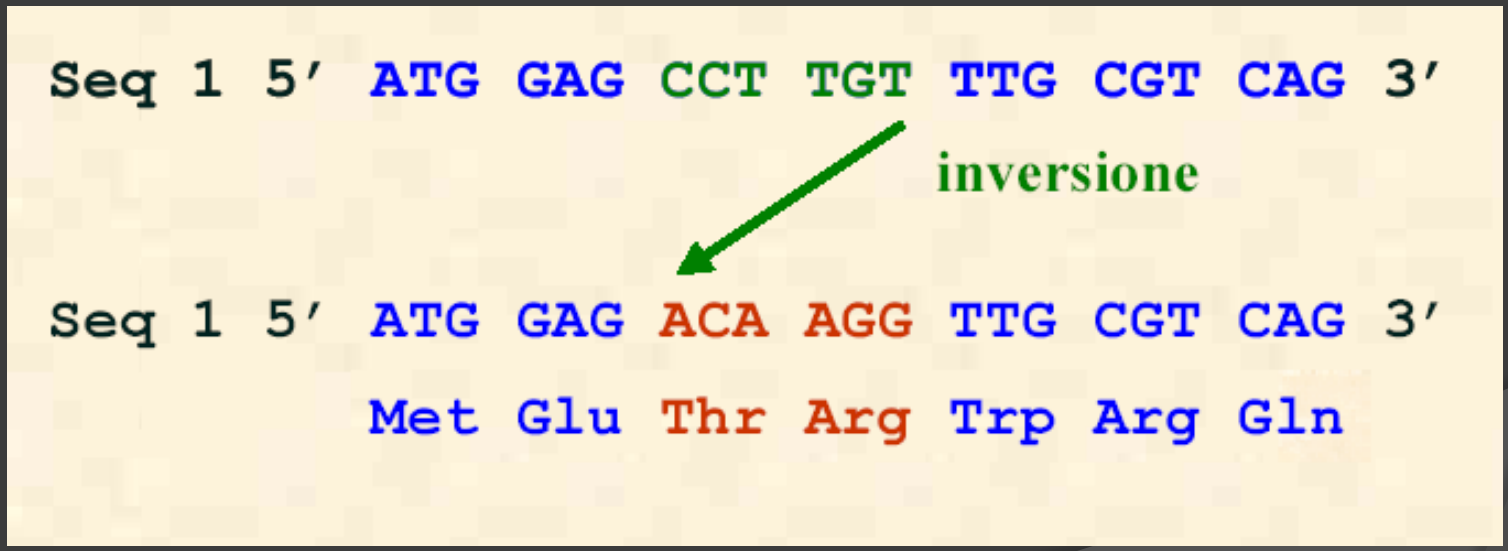
DELEZIONI


Met Glu Cys Trp Arg Gln
Seq 1 5′ ATG GAG TGT TTG CGT CAG 3′
delezione
Seq 1 5′ ATG GAG CCT TGT TTG CGT CAG 3′
delezione
Seq 1 5′ ATG GAG CCT TGT TAG 3′
Met Glu Pro Cys Ter

INSERZIONI


Met Glu Pro His Cys Trp Arg Gln
Seq 1a 5′ ATG GAG CCT CAC TGT TTG CGT CAG 3′
inserzione
Seq 1 5′ ATG GAG CCT TGT TTG CGT CAG 3′
inserzione
Seq 1b 5′ ATG GAG CCT TGA TTT GCG TCA G 3′
Met Glu Pro Ter Phe Ala Ser

INVERSIONI


Seq 1 5′ ATG GAG CCT TGT TTG CGT CAG 3′
inversione
Seq 1 5′ ATG GAG ACA AGG TTG CGT CAG 3′
Met Glu Thr Arg Trp Arg Gln

Significato dell'allineamento

Qual è la corrispondenza fra gli aminoacidi delle due sequenze che più probabilmente rispecchia l'evoluzione delle due proteine?

L'allineamento tra due sequenze biologiche è utile per scoprire informazione funzionale, strutturale ed evolutiva

Cosa vuol dire allineare due sequenze?

scrivere due sequenze orizzontalmente in modo da avere il maggior numero di simboli identici o simili in registro verticale anche introducendo intervalli (gaps – inserzioni/delezioni – *indels*)

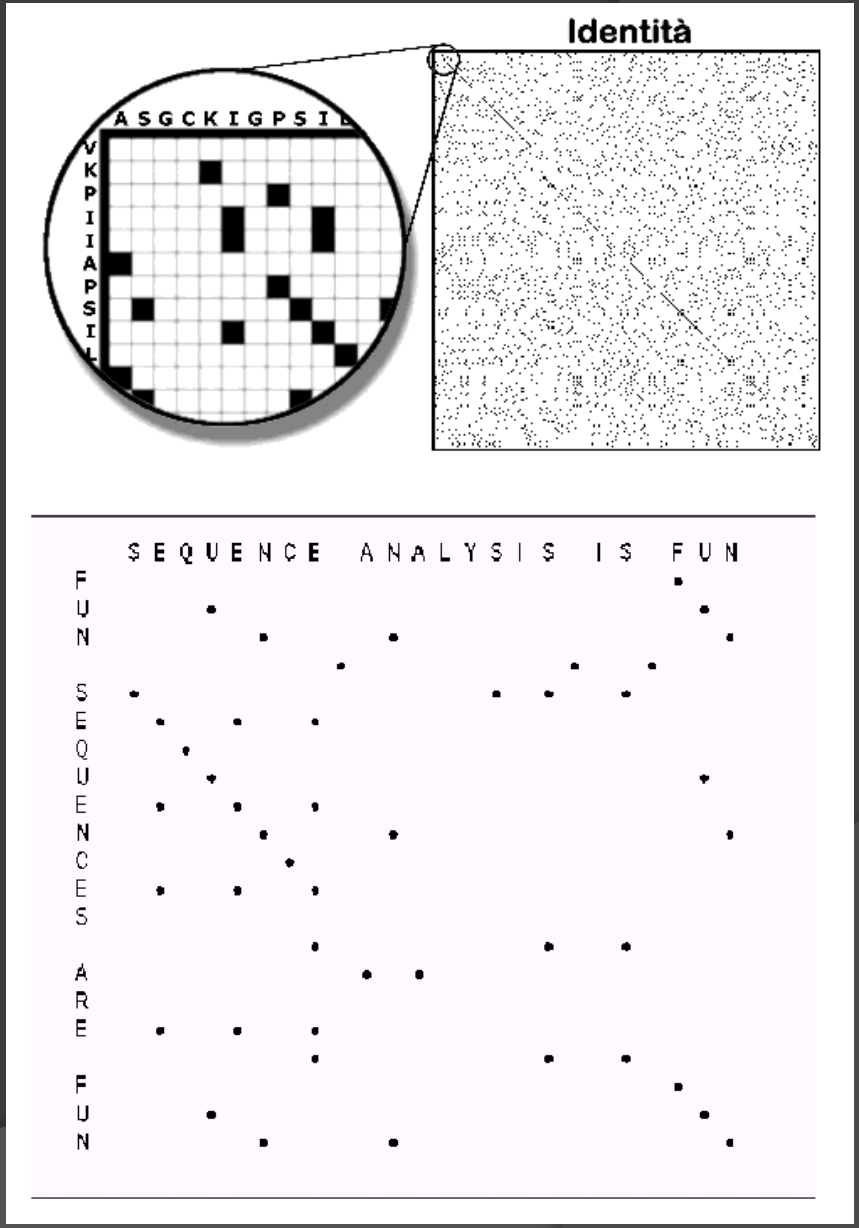
Pertanto si vogliono minimizzare le differenze o massimizzare le similarità

Metodi di allineamento

- Analisi della matrice a punti (dot matrix)
- programmazione dinamica (dynamic programming)
- metodo (Fasta, Blast)

DOT MATRIX

- Il primo semplice sistema di visualizzazione di allineamenti (1970).
- Le due sequenze da confrontare sono ai margini di una matrice. Se le due lettere corrispondenti ad una casella sono uguali allora la casella viene colorata di nero ed apparirà come un punto (dot) all'interno della matrice.
- Gli allineamenti di una certa lunghezza appaiono come segmenti diagonali e saranno immediatamente distinguibili visivamente.
- I gap appaiono come salti in diagonale.
- Le sequenze ripetute appaiono come segmenti diagonali paralleli.



LE MATRICI DI SOSTITUZIONE

- Nel caso dell'allineamento di aminoacidi è opportuno applicare dei criteri di similarità che non si limitino a verificare l'identità assoluta ma tengano conto del fatto che gli aminoacidi possano essere più o meno simili tra loro. Aminoacidi molto simili possono essere indifferentemente sostituiti in una proteina senza alcuna variazione apprezzabile nella struttura della proteina.

- Per esempio acido aspartico (D) e acido glutammico (E) sono molto simili e molto spesso nel corso dell'evoluzione prendono il posto l'uno dell'altro nelle proteine. Al contrario acido aspartico (D) e triptofano (W) sono molto diversi e non sono assolutamente interscambiabili. E' quindi ragionevole valutare differentemente la sostituzione (in generale il confronto) di D con E e di D con W.

- Ciò viene descritto in matrici quadrate di 20*20 caselle in cui si attribuisce un punteggio ad ogni possibile coppia di aminoacidi. Quanto più alto è il punteggio tanto più interscambiabili sono gli aminoacidi. Punteggi negativi penalizzano invece aminoacidi molto differenti

metodi per la valutazione del punteggio

proposta: gli allineamenti e il calcolo della similarità potrebbero essere notevolmente migliorati dall'introduzione di schemi di punteggio diversi da 0 e da 1 per l'appaiamento di residui amminoacidici

si potrebbero per esempio prevedere punteggi alti per l'identità tra coppie di residui, punteggi un po' più bassi ma >0 per residui simili dal punto di vista chimico-fisico

punteggi invece negativi (o uguali a 0) per residui diversi o molto diversi dal punto di vista chimico-fisico

ATTENZIONE

non bisogna confondere le matrici di punti con le matrici di sostituzione!

le matrici di punti sono grafici che consentono di mettere in evidenza zone di identità tra sequenze diverse. Se una sequenza è lunga m caratteri e l'altra sequenza è lunga n caratteri, la matrice di punti sarà rettangolare e di dimensione mxn

le matrici di sostituzione associano un punteggio ad ogni coppia di residui, sono matrici quadrate e simmetriche, che contengono 20x20= 400 valori, parzialmente ridondanti (il valore relativo alla coppia RK è uguale a quello della coppia KR)

ma come si calcolano i valori di una matrice di sostituzione?

Due aminoacidi i e j frequenza della loro coppia “f_{ij}”, frequenza nella sequenza di i “f_i” frequenza di j “f_j”
Rapporto “f_{ij}”/”f_i” x “f_j” misurerà quanto spesso i e j appaiono in posizioni corrispondenti non dovute al caso
log2 del rapporto valore della matrice

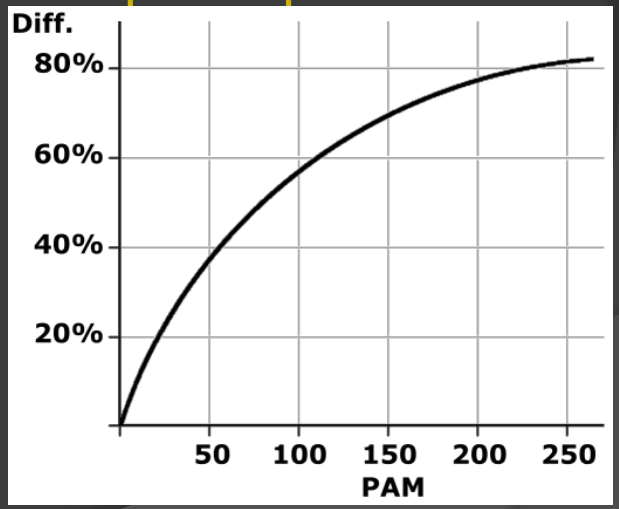
matrici di sostituzione

PAM	0	30	80	110	200	250
% identità	100	75	60	50	25	20

vediamo ora due tra le più usate matrici di sostituzione di tipo PAM: la PAM120 e la PAM250, che si utilizzano per ottimizzare allineamenti tra sequenze che abbiano circa il 50% o il 20% di identità di sequenza

PAM

- PAM1 (con i punteggi e non con le frequenze) è molto simile alla matrice Identità (valori quasi sempre 1 sulla diagonale e 0 altrove)
- PAM2 è calcolata da PAM1 ipotizzando un altro passo evolutivo e così via...
- PAMn è ottenuta da PAMn-1
- PAM100 quindi rappresenta 100 passi evolutivi in ciascuno dei quali si è avuto un 1% di sostituzioni rispetto al passo precedente.



sequenze vicine filogeneticamente

sequenze lontane filogeneticamente

PAM1 **PAM100** **PAM250**

BLOSUM

- Introdotte da Henikoff & Henikoff nel 1992.
- A differenza delle PAM generate iterativamente, queste sono invece basate su una banca dati (BLOCKS) di allineamenti multipli di segmenti proteici senza GAP.
- Il numero associato alle matrici rappresenta la percentuale di aminoacidi identici in un certo blocco

RICERCA DELLE SIMILARITA' TRA 2 SEQUENZE

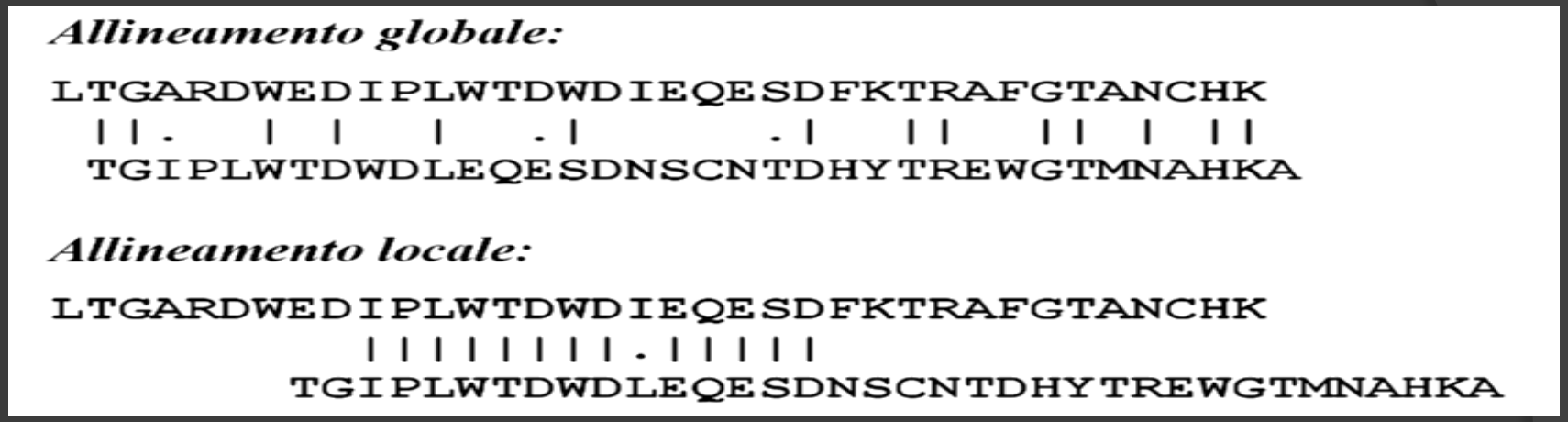
- Per determinare la similarità tra due sequenze è necessario considerare due aspetti:

- 1- ALGORITMO DI ALLINEAMENTO

- 2- CRITERIO DI SIMILARITA'

ALLINEAMENTI GLOBALI E LOCALI

Consideriamo i seguenti due differenti allineamenti delle stesse sequenze



- Nel primo caso si hanno 14 identità (evidenziate dalle linee verticali che uniscono aminoacidi uguali) e tre sostituzioni conservative (es. A-I, S-T) distribuite su tutta la lunghezza della sequenza. Nel secondo caso si hanno 13 identità ed una sostituzione conservativa su una regione di 14 aminoacidi.
- Quale dei due allineamenti è da considerarsi migliore?
- (N.B. L'allineamento globale, secondo alcuni, dovrebbe comprendere l'intera lunghezza di entrambe le sequenze; nell'esempio dovrebbero essere quindi aggiunti un gap all'inizio ed uno alla fine)

ALLINEAMENTI GLOBALI O LOCALI?

- Dal punto di vista biologico generalmente vengono privilegiati gli allineamenti locali, che riguardano regioni limitate delle proteine o di acidi nucleici.

- In Biologia Molecolare avrete sicuramente sentito parlare di domini delle proteine o anche degli acidi nucleici. Se ad esempio siamo interessati a trovare tutte le sequenze di proteine di una banca dati che contengono un certo dominio, allora sicuramente si cercheranno similarità locali.

- Gli allineamenti globali vengono applicati quando si vogliono confrontare accuratamente due sequenze in cui la similarità sia estesa per tutta la lunghezza

- N.B. Un allineamento locale non è necessariamente limitato ad una piccola regione della sequenza, ma potrebbe estendersi anche all'intera lunghezza della sequenza.

Programmazione dinamica

- Fornisce l'allineamento ottimale tra due sequenze
- semplici variazioni dell'algoritmo producono allineamento globali o locali
- l'allineamento calcolato dipende dalla scelta di alcuni parametri

Allineamento globale o locale?

1) scegliamo il miglior allineamento dal punto di vista biologico, e poi…

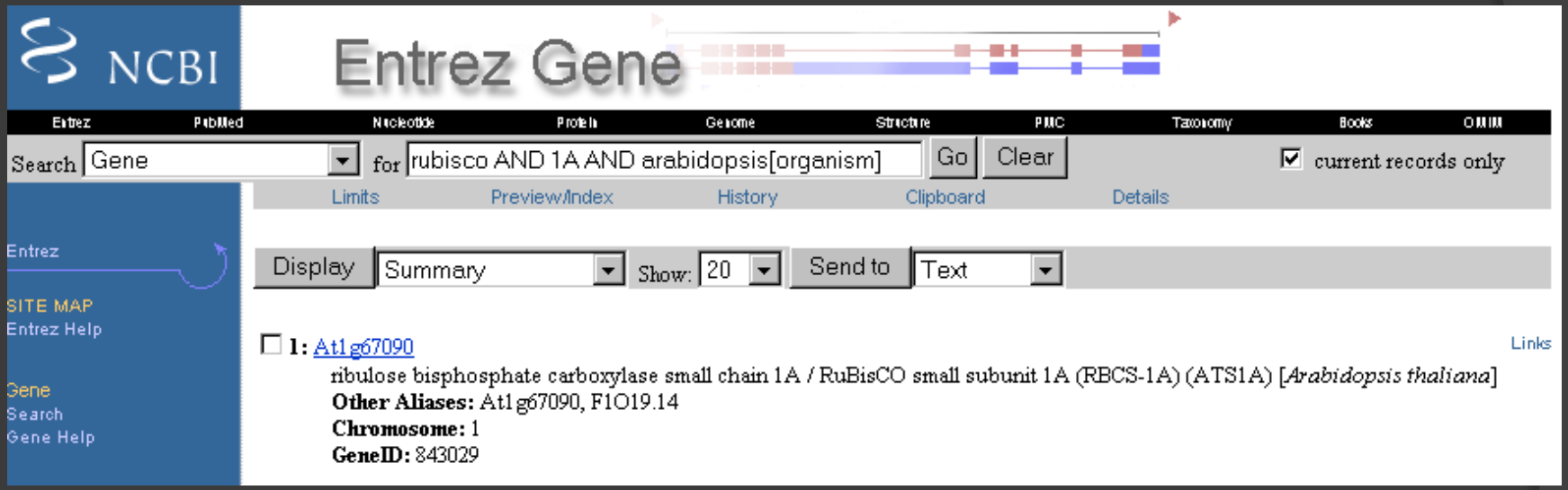
2) cerchiamo il modo di privilegiarlo dal punto di vista computazionale

spesso gli allineamenti locali hanno una migliore rispondenza con la realtà funzionale

gli allineamenti globali possono comunque essere utilizzati per confrontare accuratamente due sequenze la cui similarità sia estesa per tutta la lunghezza

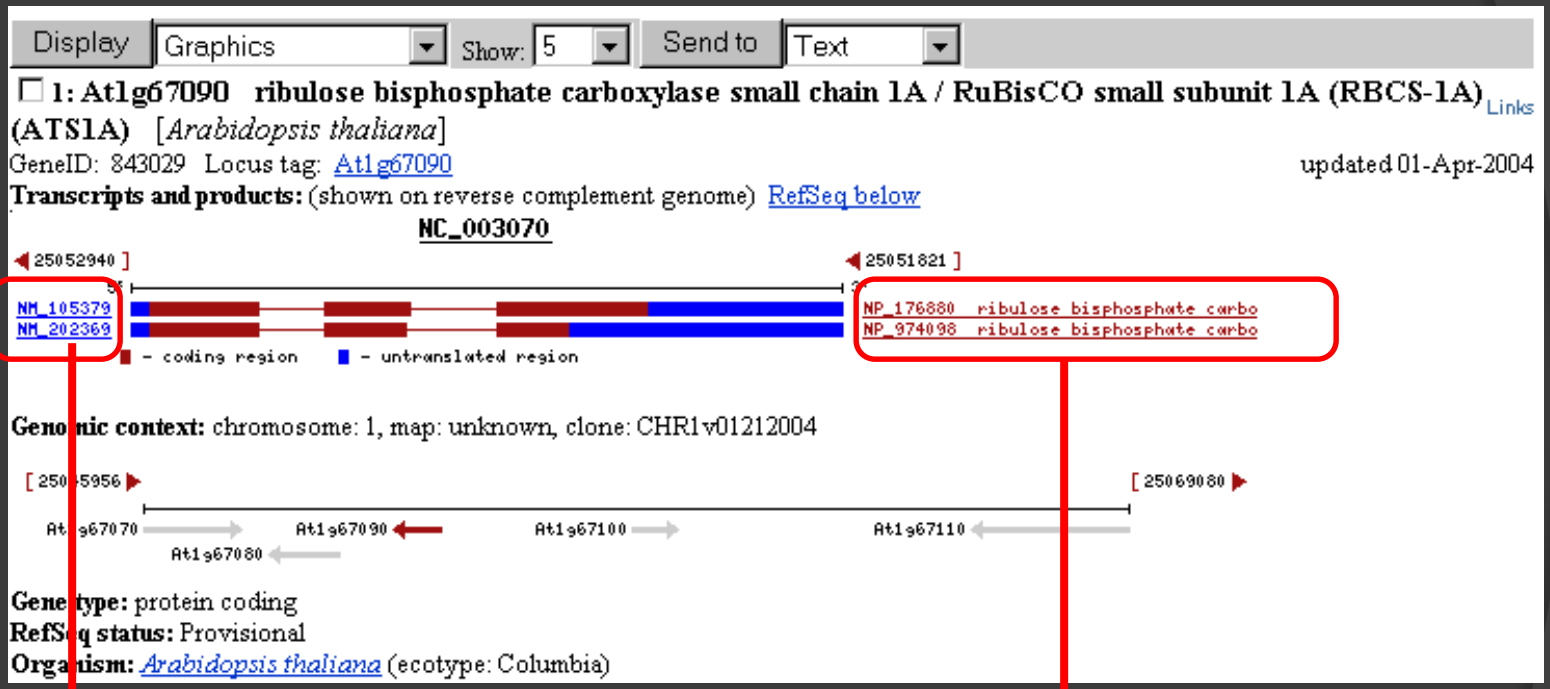
ESEMPIO

Proviamo a confontare le due diverse proteine derivanti dallo stesso gene “subunità 1A di rubisco” di Arabidopsis thaliana



ESEMPIO

Accediamo al record del gene e attraverso i link ai due diversi record delle proteine (hanno i codici NP_176880 e NP_974098)



Link ai record dei mRNA corrispondenti (nel database GENBANK)

Link ai record delle proteine corrispondenti (nel database PROTEIN)

ESEMPIO

Codici GENBANK : 15219826 e 42572015

Codici REFSEQ (database composto di sequenze “pulite e verificate” di mRNA e PROTEINE: NP_176880 e NP_974098

```
>gi|15219826|ref|NP_176880.1| ribulose bisphosphate carboxylase small chain 1A / RuBisCO small subunit 1A (RBCS-1A)
MASSMLSSATMVASPAQATMVAPFNGLKSSAAFPATRKANNDITSITSNGGRVNCMQVWPPIGKKKFETL
SYLPDLTDSELAKEVDYLIRNKWIPCVEFELEHGFVYREHGNSPGYYDGRYWTMWKLPLFGCTDSAQVLK
EVEECKKEYPNAFIRIIGFDNTRQVQCISFIAYKPPSFTG
```

180 AA

```
>gi|42572015|ref|NP_974098.1| ribulose bisphosphate carboxylase small chain 1A / RuBisCO small subunit 1A (RBCS-1A)
MASSMLSSATMVASPAQATMVAPFNGLKSSAAFPATRKANNDITSITSNGGRVNCMQVWPPIGKKKFETL
SYLPDLTDSELAKEVDYLIRNKWIPCVEFDTDLCTVSTVTHPDTMMDGTGQCGSFPCSVAPTPLKC
```

136 AA

ESEMPIO


cons
megamerger
merger
ALIGNMENT DIFFERENCES
diffseq
ALIGNMENT DOT PLOTS
dotmatcher
dotpath
dottup
polydot
ALIGNMENT GLOBAL
alignwrap
needle
ALIGNMENT LOCAL
matcher
seqmatchall
supermatcher
water
wordmatch
ALIGNMENT
>gi|15219826|ref|NP_176880.1| ribulose bisphosphate carboxylase small chain 1A / RuBisCO small subunit 1A (RBCS-1A) (ATS1A) [Arabidopsis thaliana]
MASSMLSSATMVASPAQATMVAPFNGLKSSAAFPATRKANNDITSITSNG
GRVNCMQVWPPIGKKKFETL
SYLPDLTDSELAKEVDYLIRNKWIPCVEFELEHGFVYREHGNSPGYYDGR
YWTMWKLPLFGCTDSAQVLK
Select a set of sequences.
Use one of the following three fields:
1. To access a sequence from a database, enter the USA path here:
2. Or, upload a sequence file from your local computer here:
Browse...
3. Or enter the sequence data manually here:
>gi|42572015|ref|NP_974098.1| ribulose bisphosphate carboxylase small chain 1A / RuBisCO small subunit 1A (RBCS-1A) (ATS1A) [Arabidopsis thaliana]
MASSMLSSATMVASPAQATMVAPFNGLKSSAAFPATRKANNDITSITSNG
GRVNCMQVWPPIGKKKFETL
Alignment format:
SRS format
2. SUBMIT TO NEEDLE...
run needle

ESEMPIO

```
# Aligned_sequences: 2
# 1: NP_176880.1
# 2: NP_974098.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 181
# Identity:      108/181 (59.7%)
# Similarity:    115/181 (63.5%)
# Gaps:           46/181 (25.4%)
# Score: 499.5
#
#
#=======================================

NP_176880.1         1 MASSMLSSATMVASPAQATMVAPFNGLKSSAAFPATRKANNDITSITSNG     50
                      ||||||||||||||||||||||||||||||||||||||||||||||||||
NP_974098.1         1 MASSMLSSATMVASPAQATMVAPFNGLKSSAAFPATRKANNDITSITSNG     50

NP_176880.1        51 GRVNCMQVWPPIGKKKFETLSYLPDLTDSELAKEVDYLIRNKWIPCVEFE    100
                      ||||||||||||||||||||||||||||||||||||||||||||||||||:
NP_974098.1        51 GRVNCMQVWPPIGKKKFETLSYLPDLTDSELAKEVDYLIRNKWIPCVEFD    100

NP_176880.1       101 LEHGFVYR-EHGNSPGYYDGRYWTMWKLPLFGCTDSAQVLKEVEECKKEY    149
                      .:...|.. .|.::  ..||    ..:...|.|:.:...||    |
NP_974098.1       101 TDLCTVSTVTHPDT--MMDG----TGQCGSFPCSVAPTPLK----C        136

NP_176880.1       150 PNAFIRIIGFDNTRQVQCISFIAYKPPSFTG    180

NP_974098.1       137                                 136
```

| Aminoacidi conservati
: Aminoacidi sostituiti in modo conservativo

Quando e perché l'allineamento locale?

- Confronto sequenze DNA "anonimo", per individuare sottostringhe collegate
- Individuazione subunità strutturali comuni a proteine diverse
- …

1. Ricerca di omologhe in banche dati.

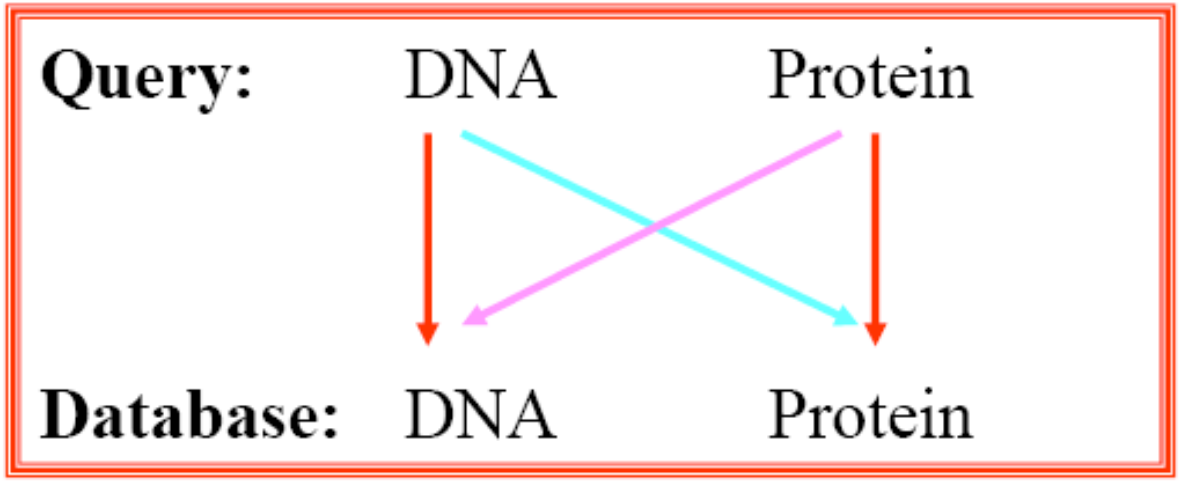
2. Programmi per la ricerca:

   FASTA

   BLAST

Ricerca di omologhe in banche dati

- Proteina *vs.* proteine
- Gene (traduzione in aa) *vs.* proteine
- Gene *vs.* geni
- Proteina *vs.* traduzione in aa di sequenze nucleotidiche (tutti i moduli)



Quando confrontiamo sequenze proteiche cerchiamo la migliore corrispondenza per 20 diversi amminoacidi

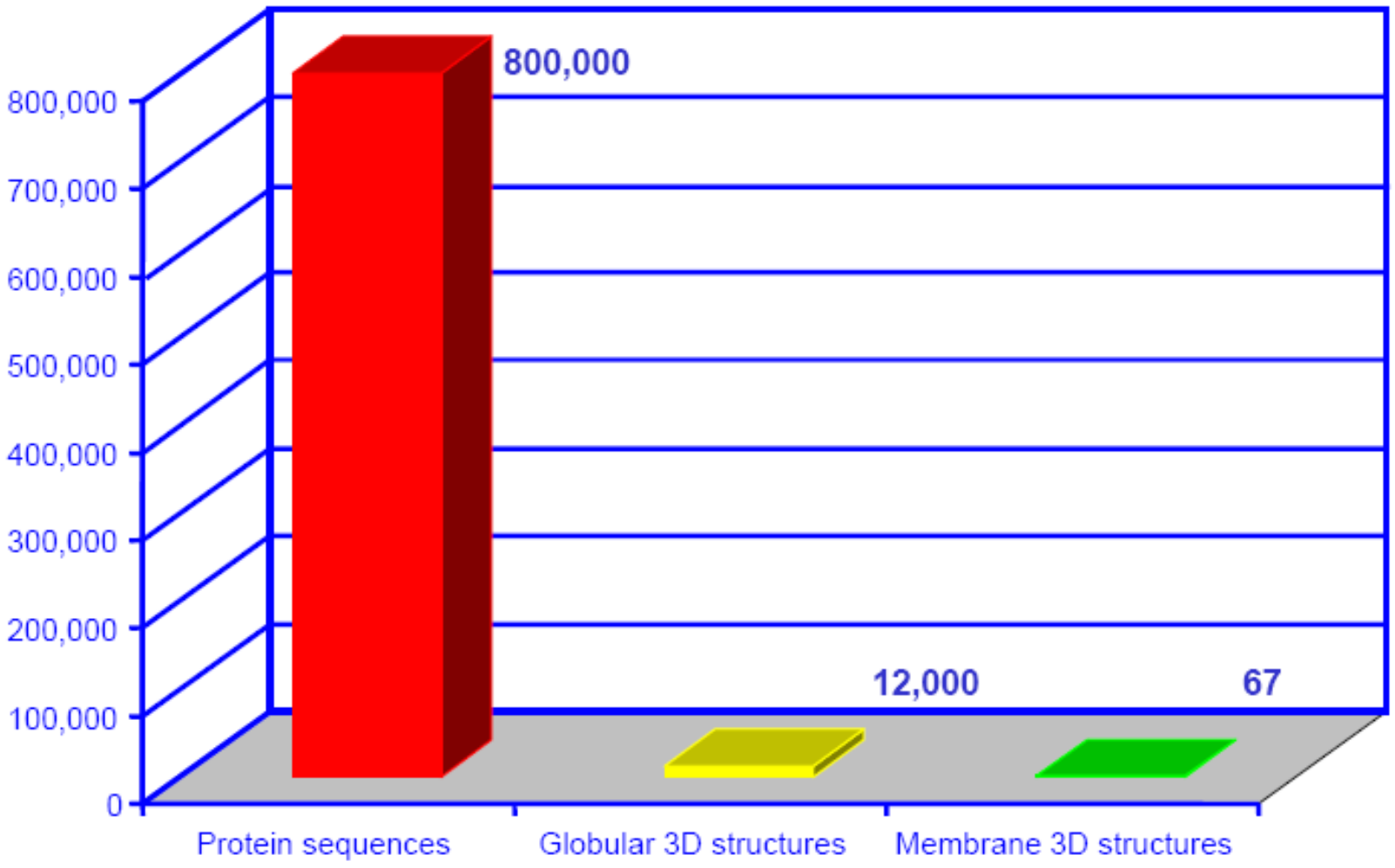
Quando confrontiamo sequenze nucleotidiche cerchiamo la migliore corrispondenza per sole 4 basi nucleotidiche

La probabilita' di trovare una buona corrispondenza (allineamento con punteggio alto) per caso è più alta per le sequenze nucleotidiche che per quelle proteiche

Inoltre, quando confrontiamo sequenze proteiche possiamo tener conto della similarita' tra i diversi amminoacidi

Quando è possibile, è preferibile confrontare sequenze proteiche !

Quante sequenze di proteine nelle banche dati ?



Come possiamo “pescare” dai databases di sequenze potenziali omologhe?

Algoritmi esatti (Smith-Waterman)
Lezione precedente

Esatto, garantisce di trovare il/i miglior allineamento/i per una coppia di sequenze.

Per 2 sequenze: A di lunghezza n and B di lunghezza m, Smith-Waterman inpiega n*m passi computazionali.

Cerchiamo omologhe della sequenza query A (n=200 aa)

Cerchiamo nel DB (10^6 sequenze di m=200 aa)

Numero di passi computazionali = 10^6 x 200 x 200 = ~10^{10}

10^3 passi al sec = 10^7 secs = 120 giorni = 4 mesi!

Necessità di algoritmi approssimati

Algoritmi esatti (Smith-Waterman)
Lezione precedente

Esatto, garantisce di trovare il/i miglior allineamento/i per una coppia di sequenze.

Per 2 sequenze: A di lunghezza n and B di lunghezza m, Smith-Waterman inpiega n*m passi computazionali.

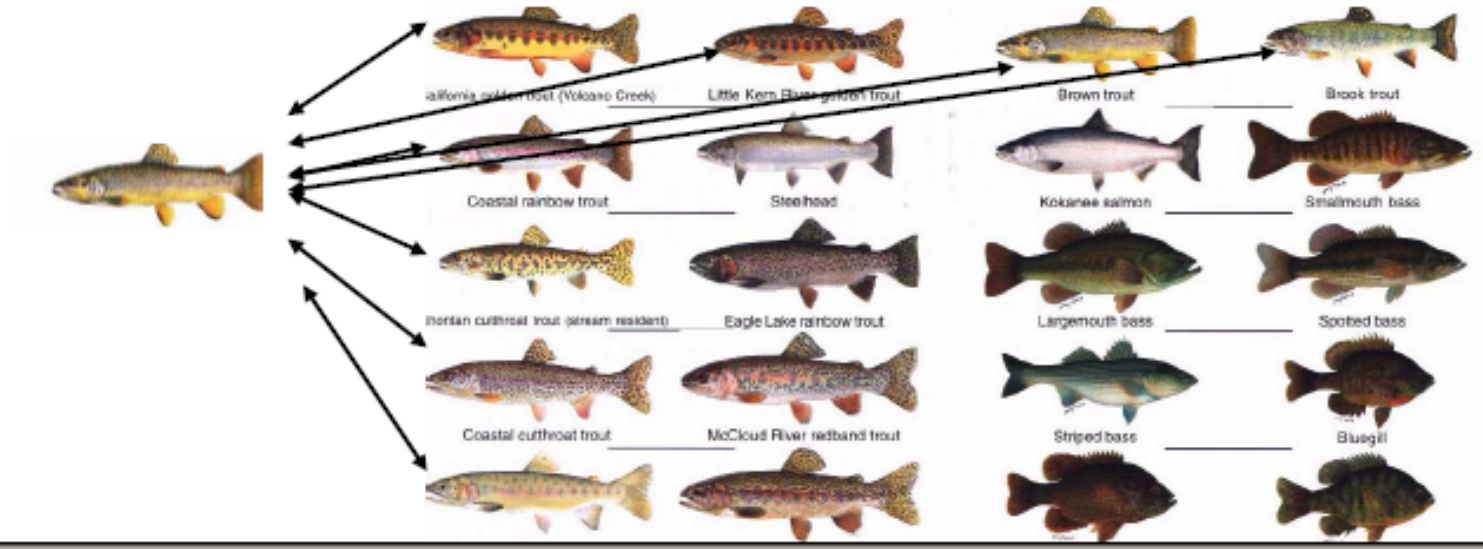
Come scartiamo gli allineamenti irrilevanti ?

Gli algoritmi euristici (BLAST, FASTA) servono a scartare la gran parte degli allineamenti irrilevanti.

Programmi quali **FASTA** e **BLAST**, partendo da una sequenza query: 

prima "pescano" dalle banche dati un sottoinsieme di sequenze che sono potenziali omologhe

poi allineano al meglio ciascuna sequenza di questo sottoinsieme alla sequenza query



Spesso, per risolvere un problema reale, si devono affrontare problemi di ottimizzazione NP-difficili, e siccome è normale che i problemi reali siano di dimensione elevata, la risoluzione di tali problemi di ottimizzazione può richiedere tempi di calcolo proibitivi. In questi casi, siccome bisogna comunque trovare una soluzione, non resta che affidarsi a procedure che non garantiscono l'ottimalità, ma che sono veloci e forniscono una soluzione spesso accettabile. Algoritmi che operano in questo modo vengono detti *algoritmi euristici* o più semplicemente euristiche.

Algoritmi EURISTICI di allineamento

Sono nati insieme alle banche dati, con lo scopo di permettere una ricerca per similarità rapida anche se meno accurata contro le migliaia di sequenze depositate.

Attualmente i programmi più utilizzati sono:

FASTA: Lipman & Pearson (1985)

BLAST: Altshul (1990)

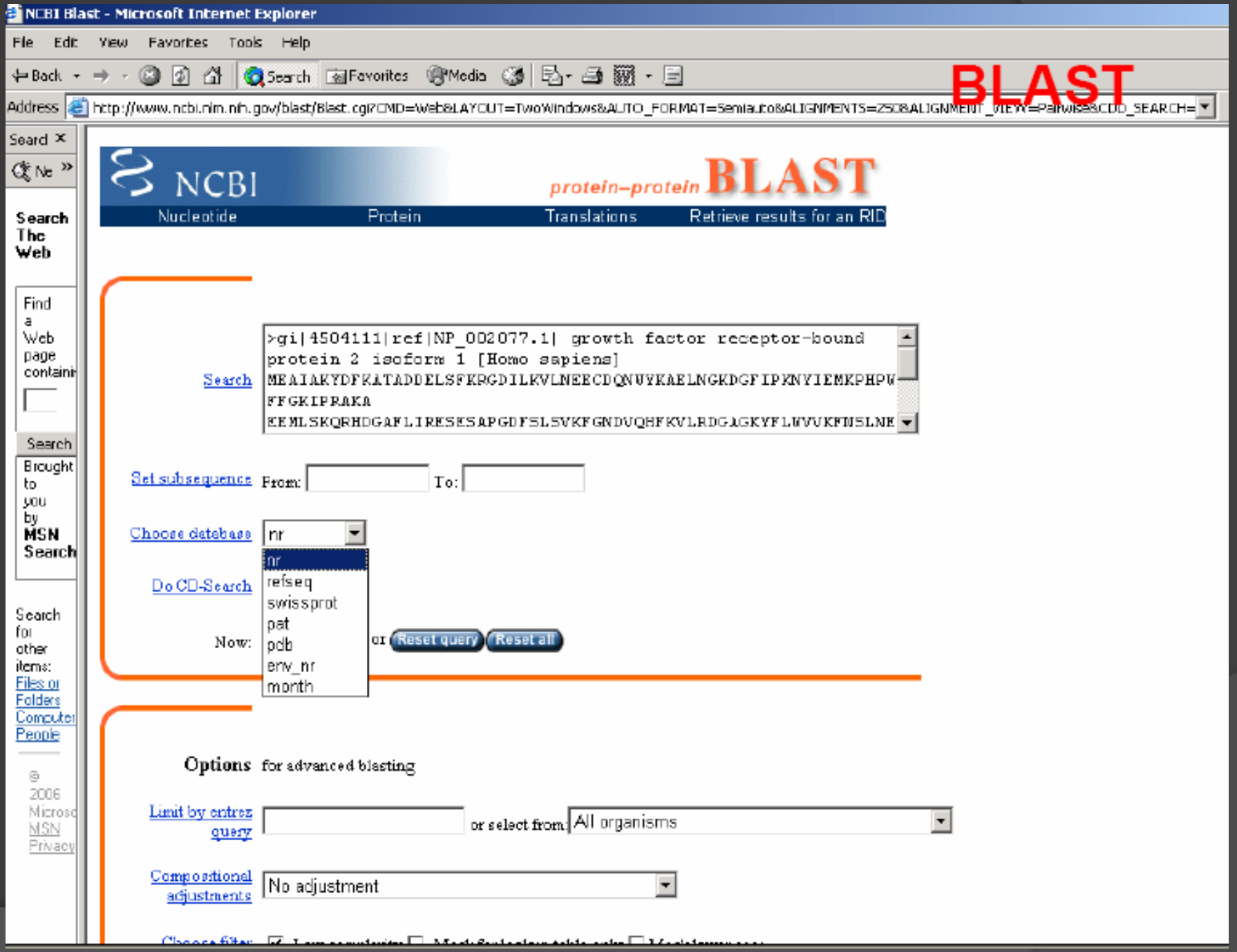
Si tratta di sequenze omologhe?
Valutazione della significatività dell'allineamento

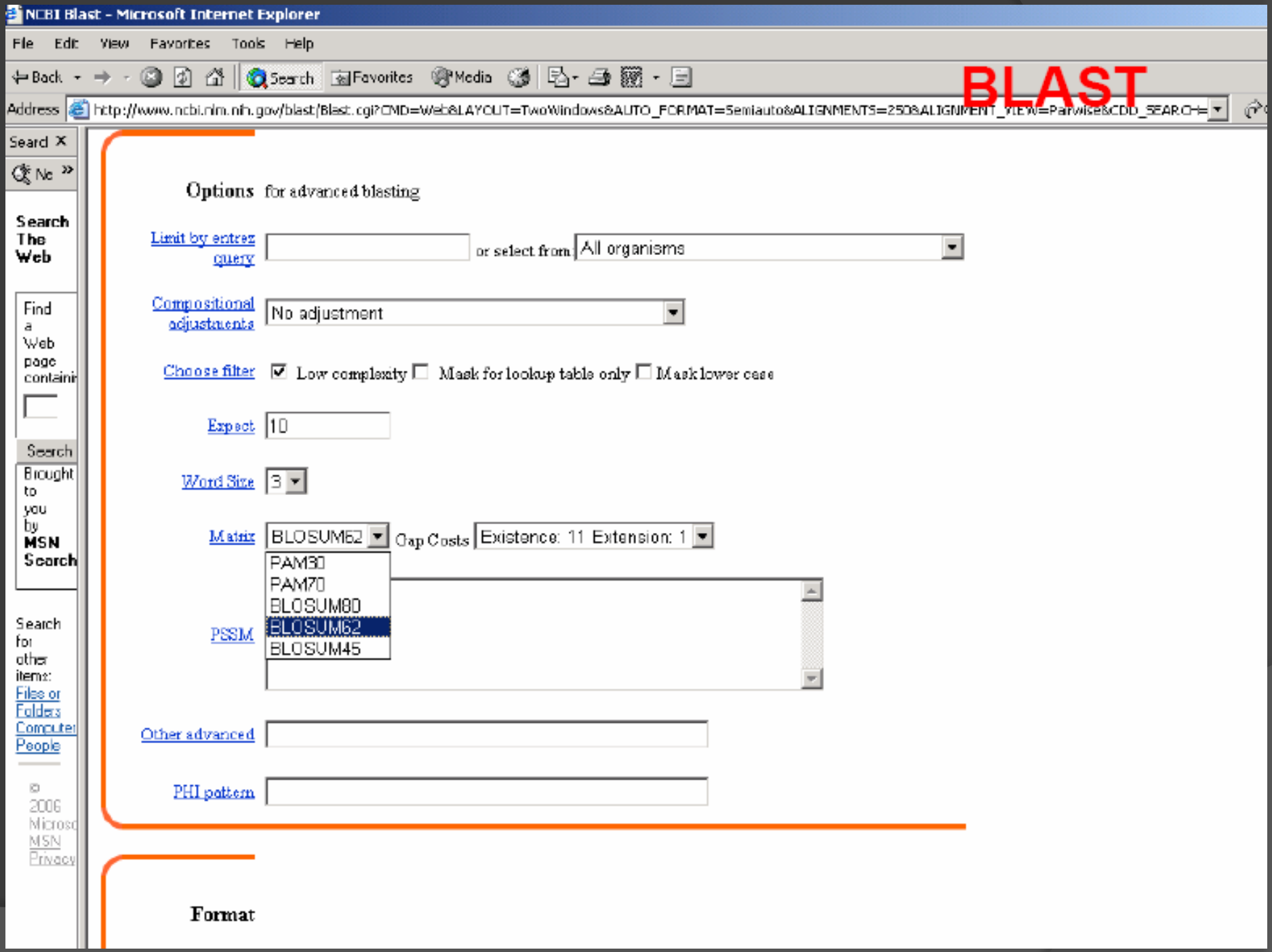
OPT grande

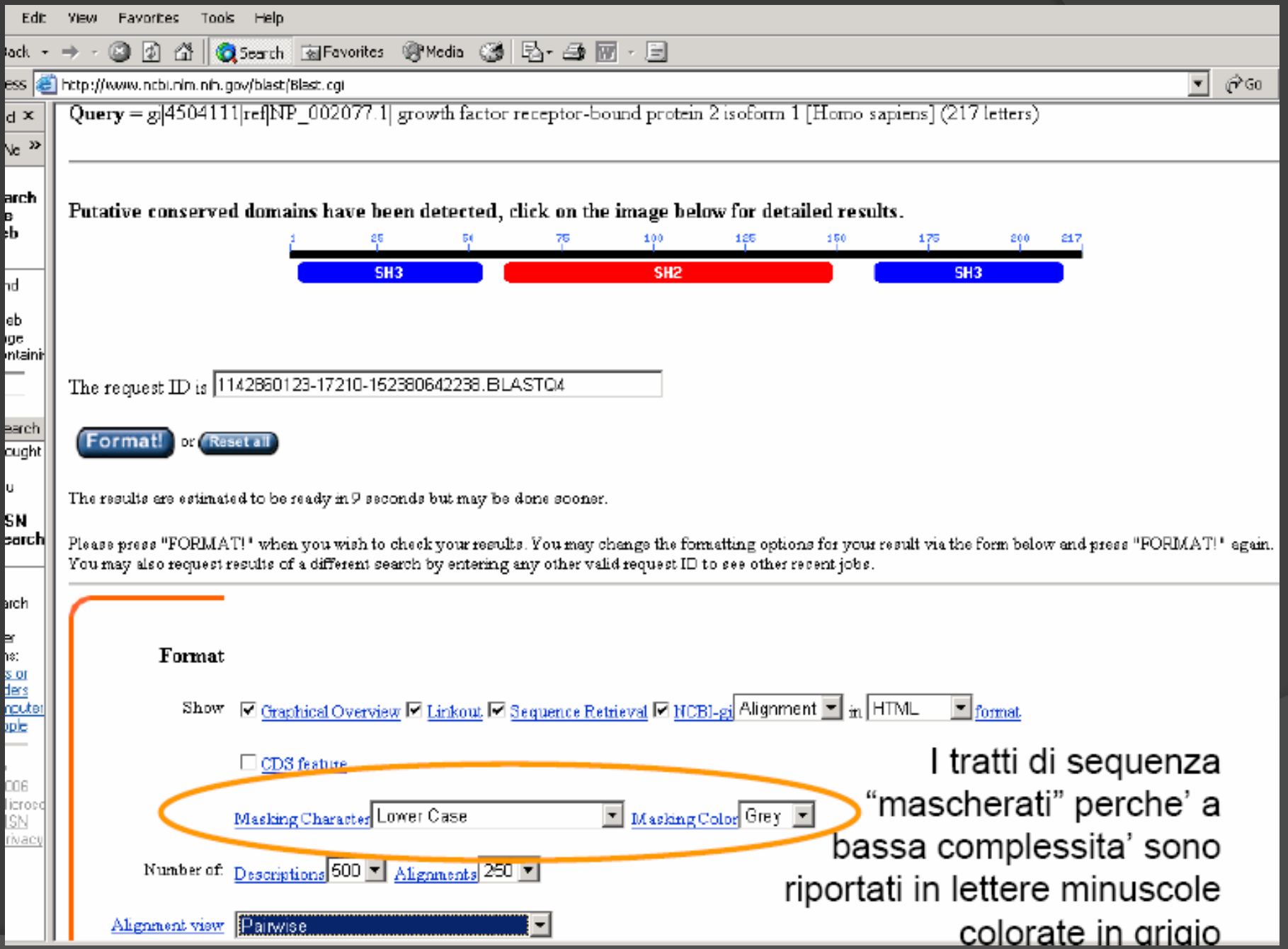
Z-score molto grande

E-value < 0.01 1.9e-110 significa 10 elevato alla –110 quindi un valore molto prossimo allo 0!

Bit-score grande


NCBI Blast - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
BLAST
Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=250&ALIGNMENT_VIEW=Pairwise&CDD_SEARCH=
Search
Search The Web
Find a Web page containi
Search
Brought to you by MSN Search
Search for other items: Files or Folders Computer People
© 2006 Microsc MSN Privacy
NCBI
protein–protein BLAST
Nucleotide Protein Translations Retrieve results for an RID
Search
>gi|4504111|ref|NP_002077.1| growth factor receptor-bound protein 2 isoform 1 [Homo sapiens]
MEAIAKYDFKATADDELSFKRGDILKVLNEECDQNWYKAELNGKDGFIPKNYIEMKPHPW
FFGKIPRAKA
EEMLSKQRHDGAFLIRESESAPGDFSLSVKFGNDVQHFKVLRDGAGKYFLWVVKFNSLNE
Set subsequence From: To:
Choose database nr
nr
refseq
swissprot
pat
pdb
env_nr
month
Do CD-Search
Now: or Reset query Reset all
Options for advanced blasting
Limit by entrez query or select from: All organisms
Compositional adjustments No adjustment
Choose filter


NCBI Blast - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
BLAST
Address http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=250&ALIGNMENT_VIEW=Pairwise&CDD_SEARCH=
Search The Web
Find a Web page containi
Search
Brought to you by MSN Search
Search for other items: Files or Folders Computer People
© 2006 Microsoft MSN Privacy
Options for advanced blasting
Limit by entrez query
or select from All organisms
Compositional adjustments
No adjustment
Choose filter
Low complexity
Mask for lookup table only
Mask lower case
Expect
10
Word Size
3
Matrix
BLOSUM62
Gap Costs
Existence: 11 Extension: 1
PAM30
PAM70
BLOSUM80
BLOSUM62
BLOSUM45
PSSM
Other advanced
PHI pattern
Format


Edit View Favorites Tools Help
Search Favorites Media
http://www.ncbi.nlm.nih.gov/blast/blast.cgi
Query = gi|4504111|ref|NP_002077.1| growth factor receptor-bound protein 2 isoform 1 [Homo sapiens] (217 letters)
Putative conserved domains have been detected, click on the image below for detailed results.
1 25 50 75 100 125 150 175 200 217
SH3
SH2
SH3
The request ID is 1142860123-17210-152380642238.BLASTQ4
Format!
or
Reset all
The results are estimated to be ready in 9 seconds but may be done sooner.
Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again.
You may also request results of a different search by entering any other valid request ID to see other recent jobs.
Format
Show
Graphical Overview
Linkout
Sequence Retrieval
NCBI-gi
Alignment
in
HTML
format
CDS feature
Masking Character
Lower Case
Masking Color
Grey
Number of:
Descriptions
500
Alignments
250
Alignment view
Pairwise
I tratti di sequenza "mascherati" perche' a bassa complessita' sono riportati in lettere minuscole colorate in grigio

Edit View Favorites Tools Help

Search Favorites Media

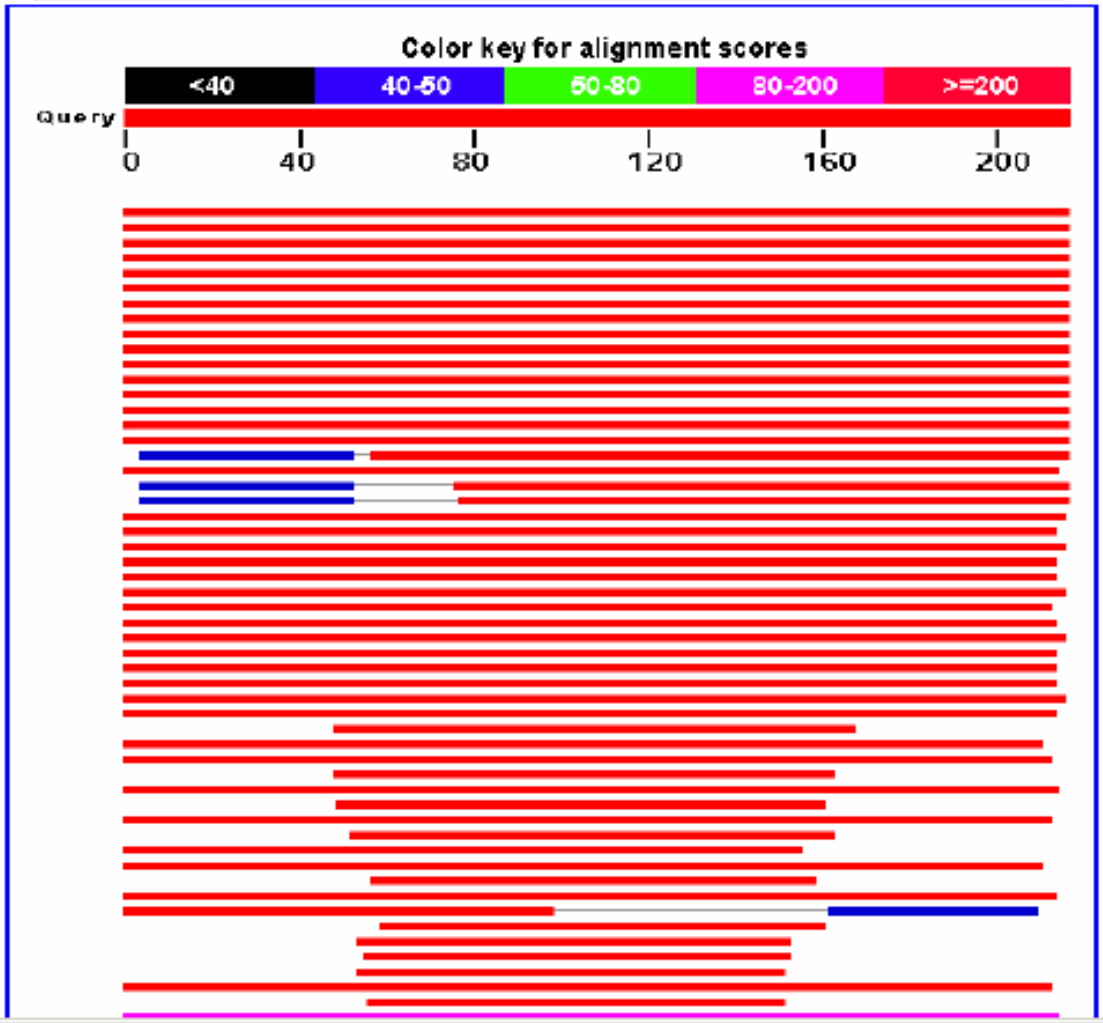
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

Go

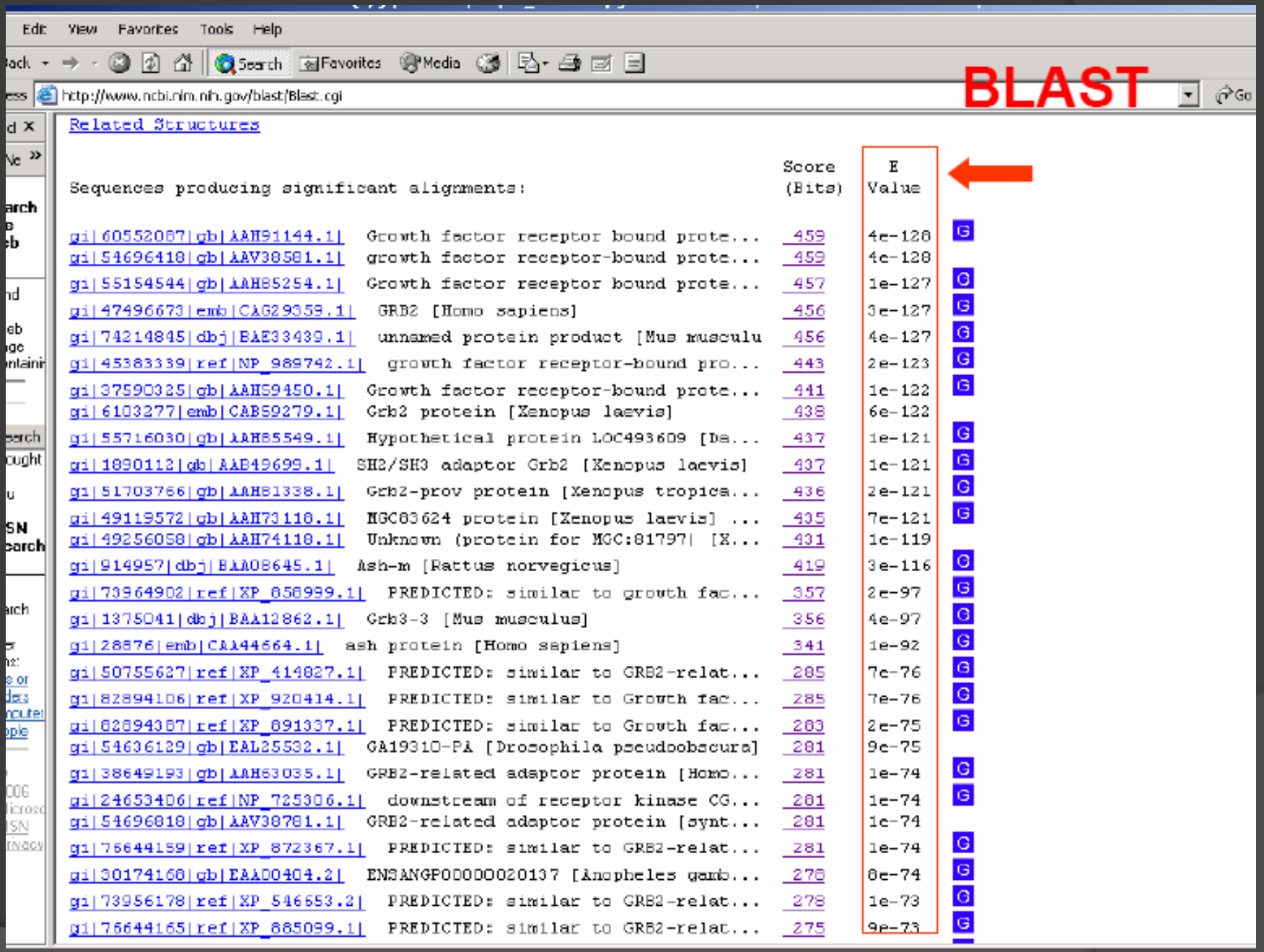
BLAST

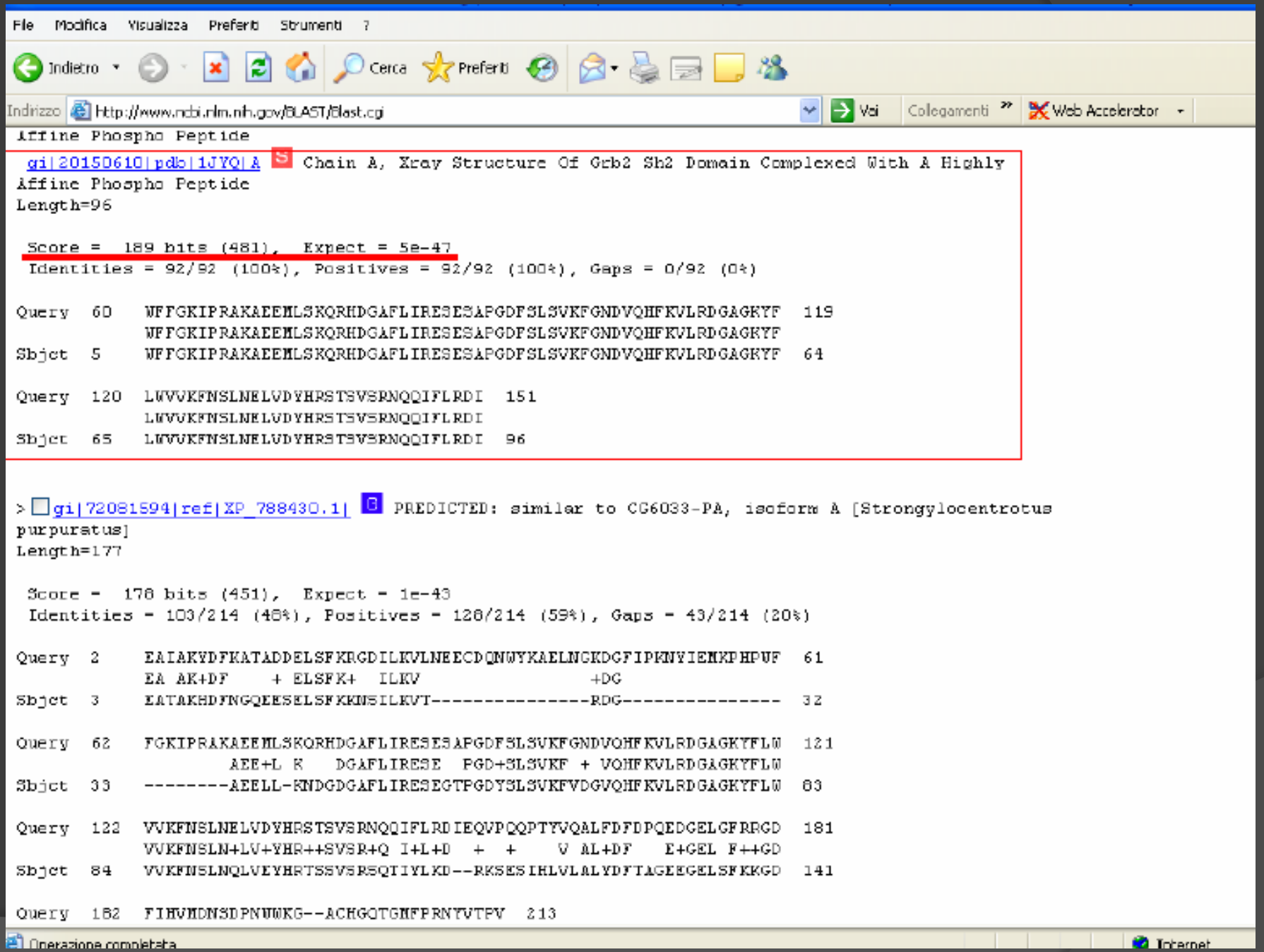
Distribution of 985 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments



Internet


Edit View Favorites Tools Help
Search Favorites Media
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
BLAST
Related Structures
Sequences producing significant alignments:
Score (Bits)
E Value
gi|60552087|gb|AAH91144.1| Growth factor receptor bound prote... 459 4e-128
gi|54696418|gb|AAV38581.1| growth factor receptor-bound prote... 459 4e-128
gi|55154544|gb|AAH85254.1| Growth factor receptor bound prote... 457 1e-127
gi|47496673|emb|CAG29359.1| GRB2 [Homo sapiens] 456 3e-127
gi|74214845|dbj|BAE33439.1| unnamed protein product [Mus musculu 456 4e-127
gi|45383339|ref|NP_989742.1| growth factor receptor-bound pro... 443 2e-123
gi|37590325|gb|AAH59450.1| Growth factor receptor-bound prote... 441 1e-122
gi|6103277|emb|CAB59279.1| Grb2 protein [Xenopus laevis] 438 6e-122
gi|55716030|gb|AAH85549.1| Hypothetical protein LOC493609 [Da... 437 1e-121
gi|1890112|gb|AAB49699.1| SH2/SH3 adaptor Grb2 [Xenopus laevis] 437 1e-121
gi|51703766|gb|AAH81338.1| Grb2-prov protein [Xenopus tropica... 436 2e-121
gi|49119572|gb|AAH73110.1| MGC83624 protein [Xenopus laevis] ... 435 7e-121
gi|49256058|gb|AAH74118.1| Unknown (protein for MGC:81797) [X... 431 1e-119
gi|914957|dbj|BAA08645.1| Ash-m [Rattus norvegicus] 419 3e-116
gi|73964902|ref|XP_858999.1| PREDICTED: similar to growth fac... 357 2e-97
gi|1375041|dbj|BAA12862.1| Grb3-3 [Mus musculus] 356 4e-97
gi|28876|emb|CAA44664.1| ash protein [Homo sapiens] 341 1e-92
gi|50755627|ref|XP_414827.1| PREDICTED: similar to GRB2-relat... 285 7e-76
gi|82894106|ref|XP_920414.1| PREDICTED: similar to Growth fac... 285 7e-76
gi|82894367|ref|XP_091337.1| PREDICTED: similar to Growth fac... 283 2e-75
gi|54636129|gb|EAL25532.1| GA19310-PA [Drosophila pseudoobscura] 281 9e-75
gi|38649193|gb|AAH63035.1| GRB2-related adaptor protein [Homo... 281 1e-74
gi|24653406|ref|NP_725306.1| downstream of receptor kinase CG... 281 1e-74
gi|54696818|gb|AAV38781.1| GRB2-related adaptor protein [synt... 281 1e-74
gi|76644159|ref|XP_872367.1| PREDICTED: similar to GRB2-relat... 281 1e-74
gi|30174168|gb|EAA00404.2| ENSANGP00000020137 [Anopheles gamb... 278 6e-74
gi|73956178|ref|XP_546653.2| PREDICTED: similar to GRB2-relat... 278 1e-73
gi|76644165|ref|XP_865099.1| PREDICTED: similar to GRB2-relat... 275 9e-73


File Modifica Visualizza Preferiti Strumenti ?
Indietro
Cerca
Preferiti
Indirizzo
http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi
Vai
Collegamenti
Web Accelerator
Affine Phospho Peptide
gi|20150610|pdb|1JYQ|A
Chain A, Xray Structure Of Grb2 Sh2 Domain Complexed With A Highly
Affine Phospho Peptide
Length=96
Score = 189 bits (481), Expect = 5e-47
Identities = 92/92 (100%), Positives = 92/92 (100%), Gaps = 0/92 (0%)
Query 60 WFFGKIPRAKAEEMLSKQRHDGAFLIRESESAPGDFSLSVKFGNDVQHFKVLRDGAGKYF 119
WFFGKIPRAKAEEMLSKQRHDGAFLIRESESAPGDFSLSVKFGNDVQHFKVLRDGAGKYF
Sbjct 5 WFFGKIPRAKAEEMLSKQRHDGAFLIRESESAPGDFSLSVKFGNDVQHFKVLRDGAGKYF 64
Query 120 LWVVKFNSLNELVDYHRSTSVSRNQQIFLRDI 151
LWVVKFNSLNELVDYHRSTSVSRNQQIFLRDI
Sbjct 65 LWVVKFNSLNELVDYHRSTSVSRNQQIFLRDI 96
gi|72081594|ref|XP_788430.1|
PREDICTED: similar to CG6033-PA, isoform A [Strongylocentrotus
purpuratus]
Length=177
Score = 178 bits (451), Expect = 1e-43
Identities = 103/214 (48%), Positives = 128/214 (59%), Gaps = 43/214 (20%)
Query 2 EAIAKYDFKATADDELSFKRGDILKVLNEECDQNWYKAELNGKDGFIPKNYIEMKPHPWF 61
EA AK+DF + ELSFK+ ILKV +DG
Sbjct 3 EATAKHDFNGQEESELSFKKNSILKVT---------------RDG--------------- 32
Query 62 FGKIPRAKAEEMLSKQRHDGAFLIRESESAPGDFSLSVKFGNDVQHFKVLRDGAGKYFLW 121
AEE+L K DGAFLIRESE PGD+SLSVKF + VQHFKVLRDGAGKYFLW
Sbjct 33 --------AEELL-KNDGDGAFLIRESEGTPGDYSLSVKFVDGVQHFKVLRDGAGKYFLW 83
Query 122 VVKFNSLNELVDYHRSTSVSRNQQIFLRDIEQVPQQPTYVQALFDFDPQEDGELGFRRGD 181
VVKFNSLN+LV+YHR++SVSR+Q I+L+D + + V AL+DF E+GEL F++GD
Sbjct 84 VVKFNSLNQLVEYHRTSSVSRSQTIYLKD--RKSESIHLVLALYDFTAGEEGELSFKKGD 141
Query 182 FIHVMDNSDPNWWKG--ACHGQTGMFPRNYVTPV 213
Operazione completata
Internet

Per ricerche in banche dati nucleotidiche, l'indicizzazione in w-mers ha poca rilevanza.

Inoltre il valore w di default di Blast per i nucleotidi è 11, il che lo porta a non riconoscere sequenze che condividano in modo esatto meno di 11 basi, è questo è un limite grosso.

Fasta è molto più tollerante per sequenze che presentano gaps, visto che già nelle prime fasi prevede il loro inserimento, mentre Blast li inserisce solo in fase di allungamento.

$\Rightarrow$ FASTA è più adatto a ricerche in banche dati nucleotidiche

$\Rightarrow$ BLAST è più adatto a ricerche in banche dati proteiche

Anche se questa regola è un po' troppo arbitraria...

Perché l'allineamento multiplo?

Il modo più efficace per decifrare un testo scritto in lingua sconosciuta è quello di confrontarlo con testi equivalenti scritti in linguaggi noti (Stele di Rosetta 196 a.C)

Allineamento multiplo: confrontare più sequenze ricerca di omologhe

Omologia (carattere qualitativo): geni o proteine che hanno un comune progenitore

Ortologhi: divergenza di sequenza dovuta ad un evento di speciazione (Beta-globina ratto e umana)

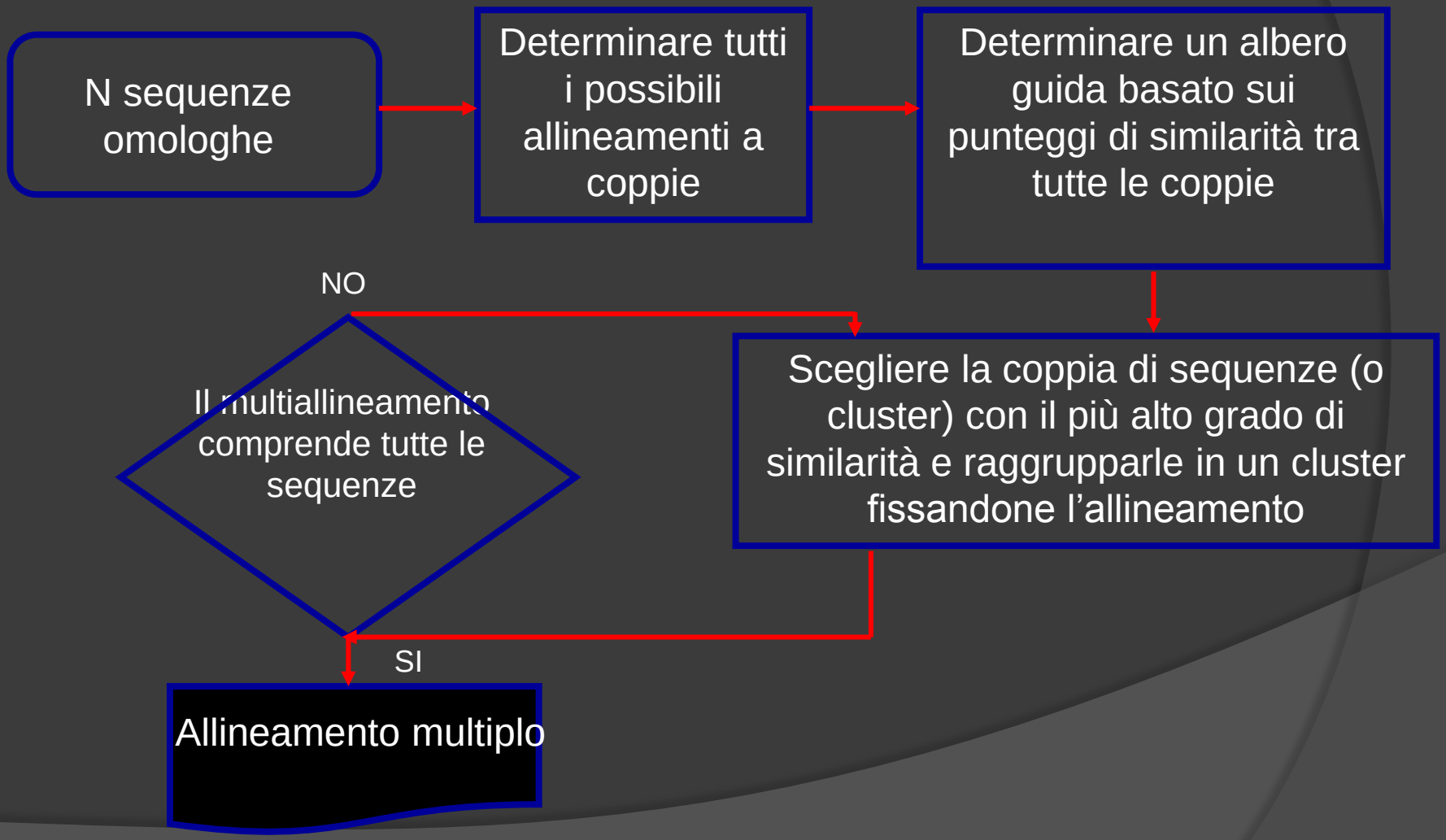
Paraloghi: divergenza in caso di duplicazione genica (alfa e beta globine umana)

Per potere affermare che due sequenze siano omologhe ci si affida al criterio di similarità. Se due sequenze possiedono un elevato grado di similarità sarà improbabile che cio sia dovuto esclusivamente al caso.

Non si può escludere il contrario

Organismi molto lontani evolutivamente

Similarità per convergenza adattativa


N sequenze omologhe
Determinare tutti i possibili allineamenti a coppie
Determinare un albero guida basato sui punteggi di similarità tra tutte le coppie
Scegliere la coppia di sequenze (o cluster) con il più alto grado di similarità e raggrupparle in un cluster fissandone l'allineamento
Il multiallineamento comprende tutte le sequenze
NO
SI
Allineamento multiplo

Pôle BioInformatique Lyonnais
Network Protein Sequence Analysis

NPS@ is the IBCP contribution to PBIL in Lyon, France

Wednesday, December 28th 2005: NPS@ server will not be available from January 9th to 13th. We apologize for any inconveniences.

Monday, September 26th 2005 : fixed secondary structure prediction insertion in Multalin. (see news)
When sending automatic requests on NPS@, please use HTTP POST method not GET.

CLUSTALW

[Abstract] [NPS@ help] [Original server]

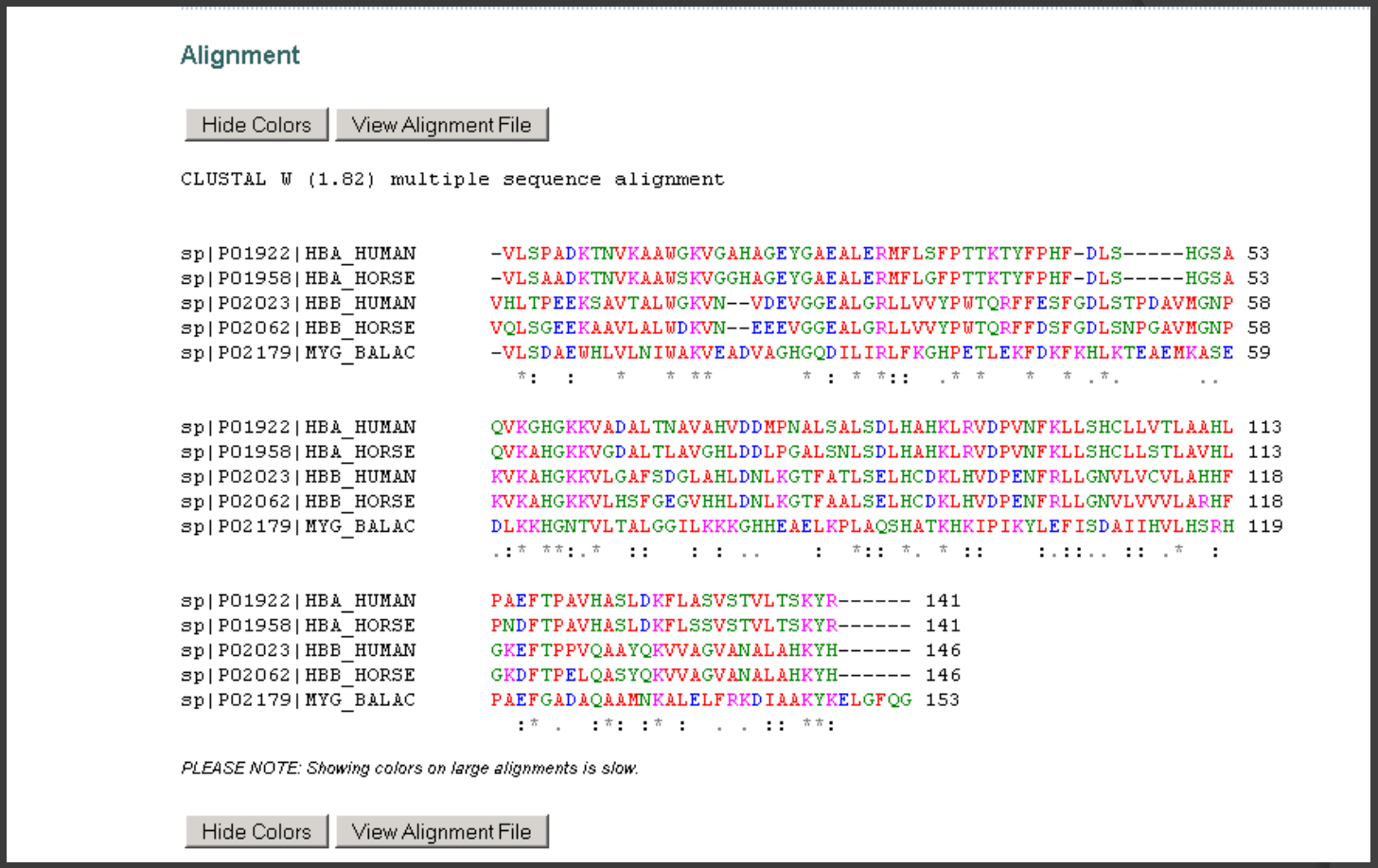
Paste a protein sequence databank in Pearson/Fasta format below : help

```
>sp|P09345|HEMA_IACKS Hemagglutinin precursor [Contains:
Hemagglutinin HA1 chain; Hemagglutinin HA2 chain] -
Influenza A virus (strain A/Chicken/Scotland/59 H5N1).
MERIVLLLAIVSLVKSDQICIGYHANKSTKQVDTIMEKNVTVTHAQDILERTHNGKLCSL
NGVKPLILRDCSVAGWLLGNPMCDEFLNLPEWLYIVEKDNPINSLCYPGDFNDYEELKYL
LSSTNHFEKIRIIPRSSWSNHDASSGVSSACPYIGRSSFLRNVVWLIKKNNTYPTIKRSY
NNTNQEDLLILWGIHHPNDAAEQTKLYQNPTTYVSVGTSTLNQRSIPEIATRPKVNGQSG
RMEFFWTILKPNDAINFESNGNFIAPRYAYKIVKKGDSAIMKSGLAYGNCDTKCQTPVGE
INSSMPFHNIHPHTIGECPKYVKSDRLVLATGLRNVPQRKKRGLFGAIAGFIEGGWQGMV
```

All sequence names must be different !

SUBMIT CLEAR

Output width : 60



Alignment

Hide Colors | View Alignment File

```
CLUSTAL W (1.82) multiple sequence alignment


sp|P01922|HBA_HUMAN      -VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSA 53
sp|P01958|HBA_HORSE      -VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-----HGSA 53
sp|P02023|HBB_HUMAN      VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP 58
sp|P02062|HBB_HORSE      VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSNPGAVMGNP 58
sp|P02179|MYG_BALAC      -VLSDAEWHLVLNIWAKVEADVAGHGQDILIRLFKGHPETLEKFDKFKHLKTEAEMKASE 59
                           *:  :   *   * **       * : * *::  .* *   *  * .*.      ..

sp|P01922|HBA_HUMAN      QVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHL 113
sp|P01958|HBA_HORSE      QVKAHGKKVGDALTLAVGHLDDLPGALSNLSDLHAHKLRVDPVNFKLLSHCLLSTLAVHL 113
sp|P02023|HBB_HUMAN      KVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHF 118
sp|P02062|HBB_HORSE      KVKAHGKKVLHSFGEGVHHLDNLKGTFAALSELHCDKLHVDPENFRLLGNVLVVVLARHF 118
sp|P02179|MYG_BALAC      DLKKHGNTVLTALGGILKKKGHHEAELKPLAQSHATKHKIPIKYLEFISDAIIHVLHSRH 119
                         .:* **:.*  ::   : : ..    :  *:: *. * ::    :.::.. :: .*  :

sp|P01922|HBA_HUMAN      PAEFTPAVHASLDKFLASVSTVLTSKYR------ 141
sp|P01958|HBA_HORSE      PNDFTPAVHASLDKFLSSVSTVLTSKYR------ 141
sp|P02023|HBB_HUMAN      GKEFTPPVQAAYQKVVAGVANALAHKYH------ 146
sp|P02062|HBB_HORSE      GKDFTPELQASYQKVVAGVANALAHKYH------ 146
sp|P02179|MYG_BALAC      PAEFGADAQAAMNKALELFRKDIAAKYKELGFQG 153
                           :* .  :*: :* :  . . :: **:
```

PLEASE NOTE: Showing colors on large alignments is slow.

Hide Colors | View Alignment File

Alignment data :

Alignment length : 575

Identity (*) : 361 is 62.78 %

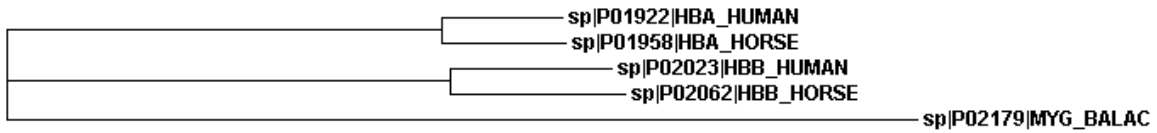
Strongly similar (:) : 86 is 14.96 %

Weakly similar (.) : 32 is 5.57 %

Different : 96 is 16.70 %

```
(
(
sp|P01922|HBA_HUMAN:0.05940,
sp|P01958|HBA_HORSE:0.06117)
:0.21271,
(
sp|P02023|HBB_HUMAN:0.07873,
sp|P02062|HBB_HORSE:0.08566)
:0.21751,
sp|P02179|MYG_BALAC:0.44687);
```

Phylogram



Show as Cladogram Tree | Show Distances | View DND File

Right-click on the above tree to see display options.

Problems printing? Read how to print a Phylogram or Cladogram.

La bonta dell'allineamento dipende dall'algoritmo ma anche dalla scelta delle sequenze.

L'allineamento risulta problematico anche quando le sequenze sono di diversa lunghezza.

ClustalW che ha la peculiarità di allineare una nuova sequenza ad un allineamento precedente.

Identità amminoacidica superiore 50%
Identità nucleotidica superiore 70%

Quando non ricorrano tali condizioni, l'allineamento richiede una ulteriore fase di aggiustamento, che tenga conto di informazioni addizionali quali:

- Conservazioni di siti funzionali e catalitici noti a priori
- Predizioni di struttura secondaria
- Predizioni di struttura terziaria

La struttura di PSI-BLAST

PSI-BLAST: usa una ricerca iterativa per cui le sequenze trovate a ogni ciclo sono
Usate per costruire un modello di punteggio per il ciclo successivo.

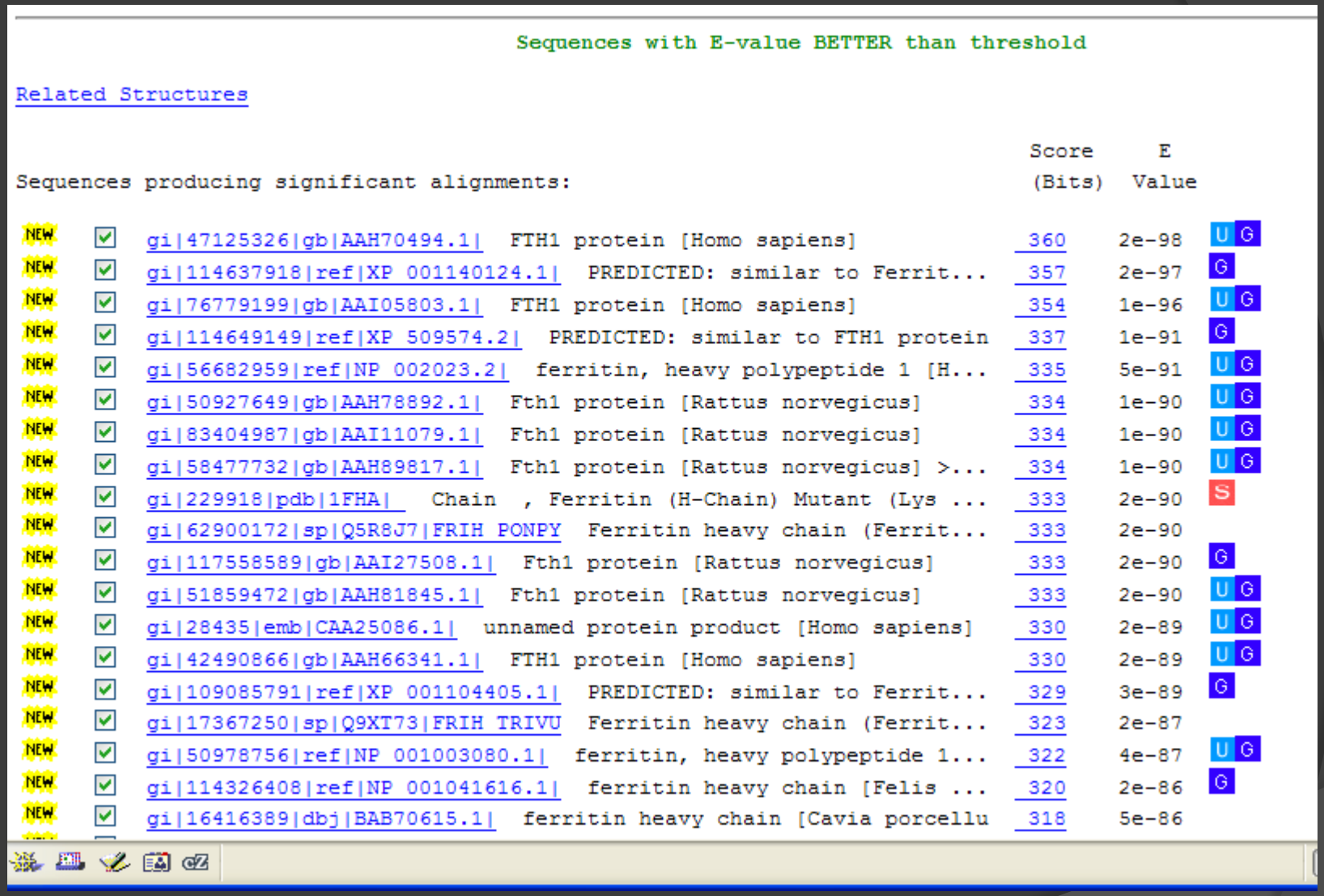
Tutta l'informazione contenuta nel multiallineamento può essere contenuta nel suo "profilo" Tra le principali applicazioni dei profili troviamo la ricerca di sequenze omologhe molto divergenti.

1° iterazione normale ricerca di similarità con l'algoritmo BLAST

2° iterazione individuato il profilo, seconda ricerca

Risultati inaccurati se si inseriscono nell'allineamento per il profilo sequenza intruse

ESEMPIO: ferritina umana

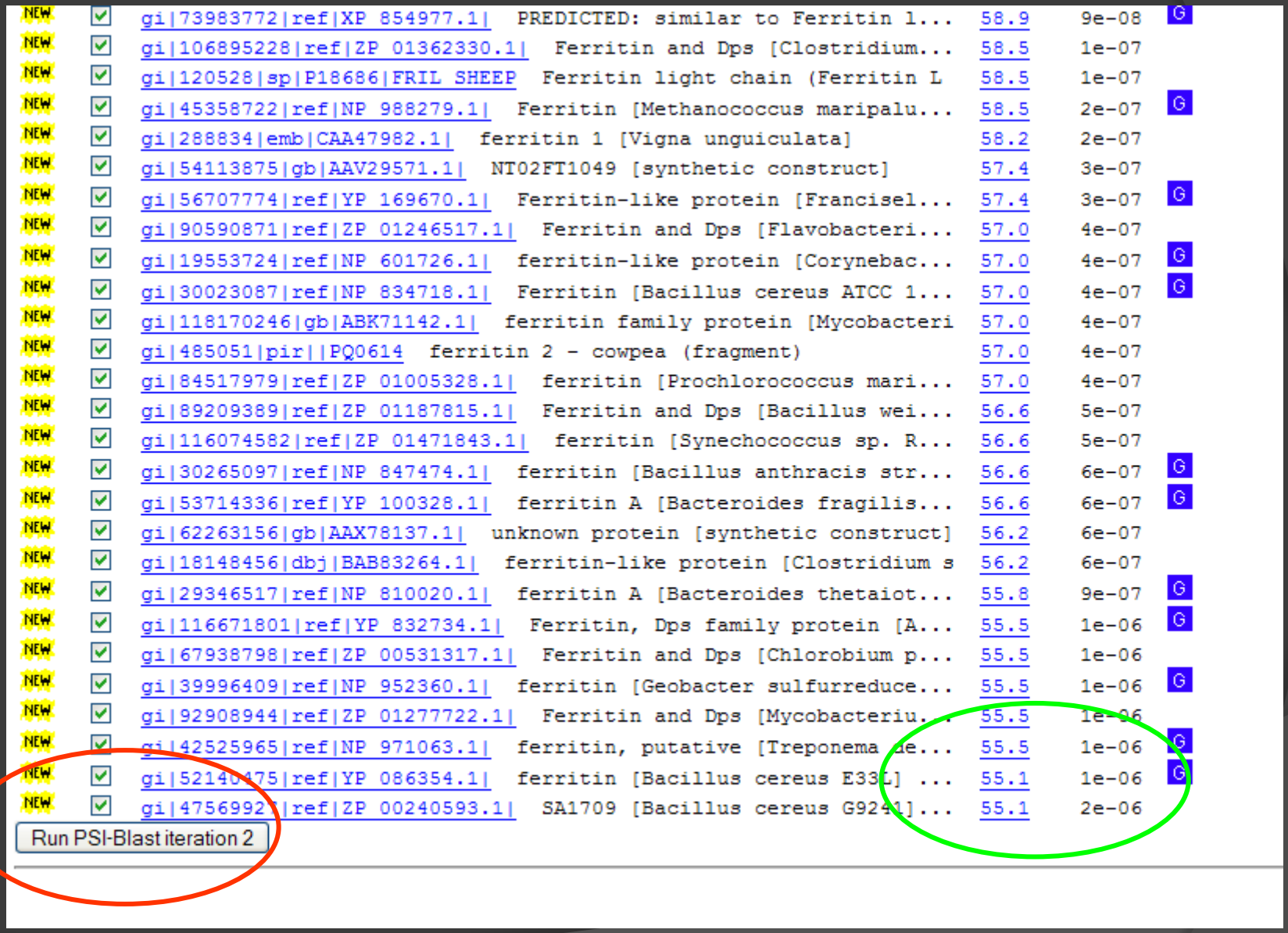


Sequences with E-value BETTER than threshold

Related Structures

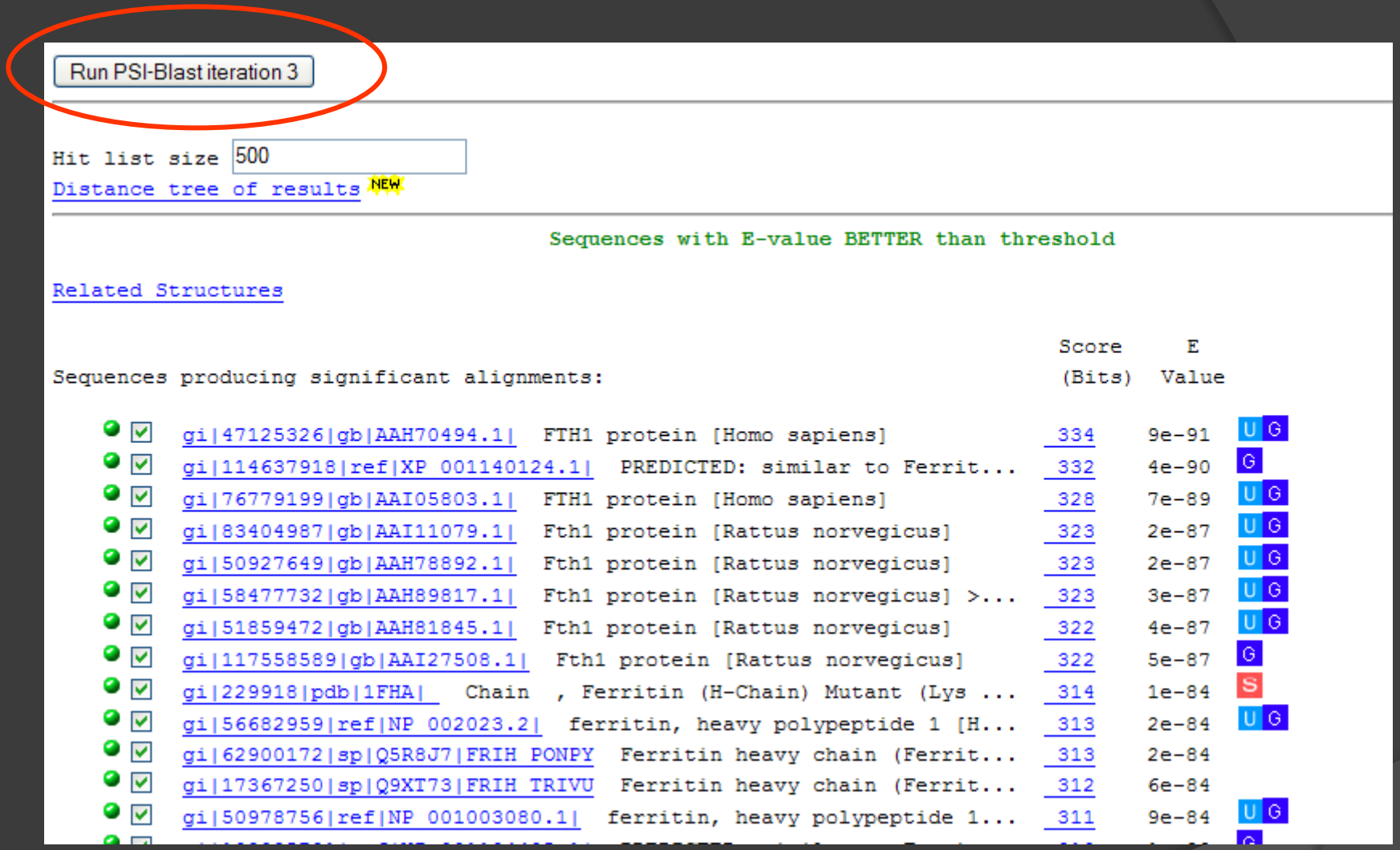
Sequences producing significant alignments:		Score (Bits)	E Value	
NEW	gi|47125326|gb|AAH70494.1| FTH1 protein [Homo sapiens]	360	2e-98	U G
NEW	gi|114637918|ref|XP_001140124.1| PREDICTED: similar to Ferrit...	357	2e-97	G
NEW	gi|76779199|gb|AAI05803.1| FTH1 protein [Homo sapiens]	354	1e-96	U G
NEW	gi|114649149|ref|XP_509574.2| PREDICTED: similar to FTH1 protein	337	1e-91	G
NEW	gi|56682959|ref|NP_002023.2| ferritin, heavy polypeptide 1 [H...	335	5e-91	U G
NEW	gi|50927649|gb|AAH78892.1| Fth1 protein [Rattus norvegicus]	334	1e-90	U G
NEW	gi|83404987|gb|AAI11079.1| Fth1 protein [Rattus norvegicus]	334	1e-90	U G
NEW	gi|58477732|gb|AAH89817.1| Fth1 protein [Rattus norvegicus] >...	334	1e-90	U G
NEW	gi|229918|pdb|1FHA| Chain , Ferritin (H-Chain) Mutant (Lys ...	333	2e-90	S
NEW	gi|62900172|sp|Q5R8J7|FRIH_PONPY Ferritin heavy chain (Ferrit...	333	2e-90	
NEW	gi|117558589|gb|AAI27508.1| Fth1 protein [Rattus norvegicus]	333	2e-90	G
NEW	gi|51859472|gb|AAH81845.1| Fth1 protein [Rattus norvegicus]	333	2e-90	U G
NEW	gi|28435|emb|CAA25086.1| unnamed protein product [Homo sapiens]	330	2e-89	U G
NEW	gi|42490866|gb|AAH66341.1| FTH1 protein [Homo sapiens]	330	2e-89	U G
NEW	gi|109085791|ref|XP_001104405.1| PREDICTED: similar to Ferrit...	329	3e-89	G
NEW	gi|17367250|sp|Q9XT73|FRIH_TRIVU Ferritin heavy chain (Ferrit...	323	2e-87	
NEW	gi|50978756|ref|NP_001003080.1| ferritin, heavy polypeptide 1...	322	4e-87	U G
NEW	gi|114326408|ref|NP_001041616.1| ferritin heavy chain [Felis ...	320	2e-86	G
NEW	gi|16416389|dbj|BAB70615.1| ferritin heavy chain [Cavia porcellu	318	5e-86	

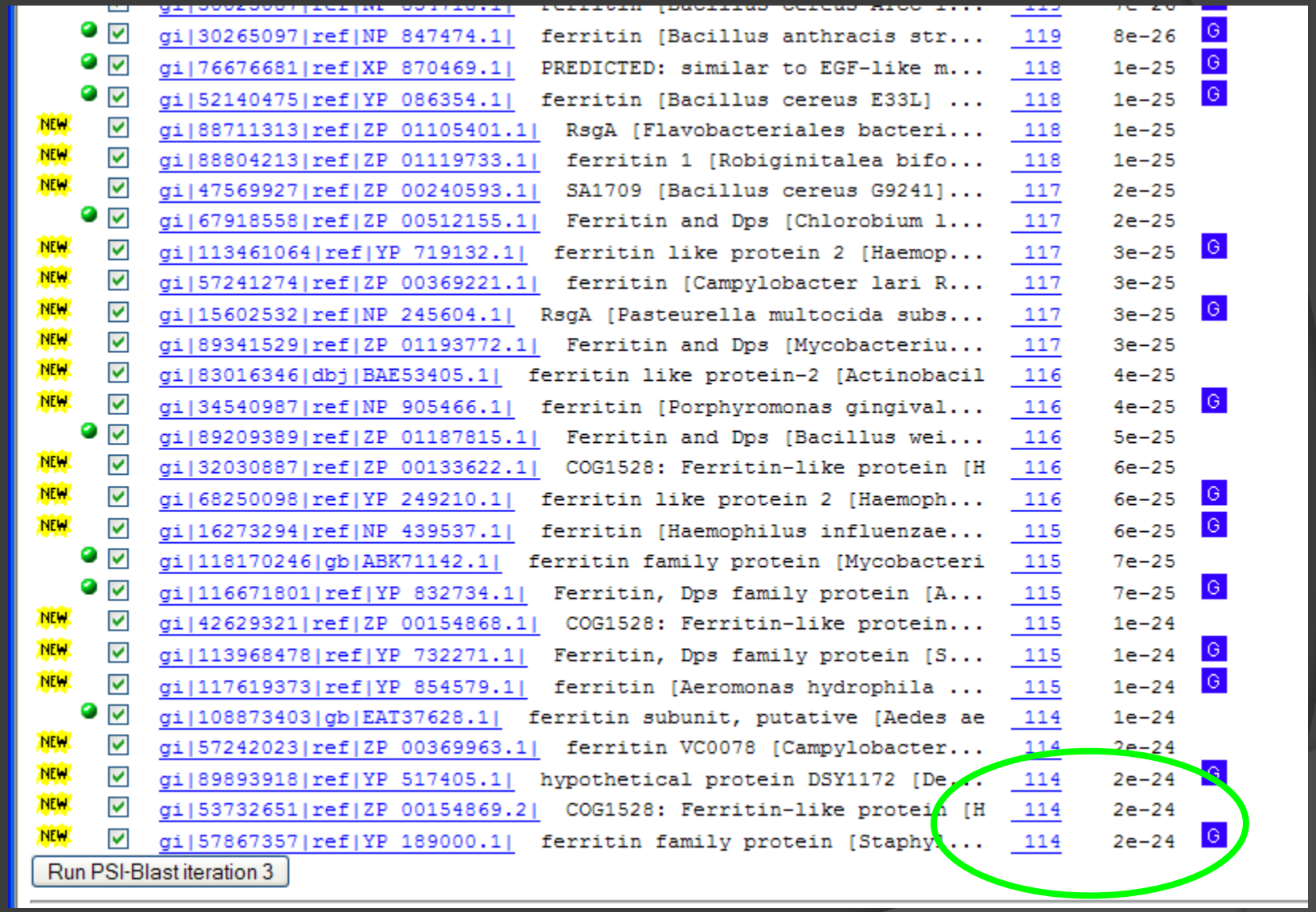
XP_043424 Ferritina umana



NEW	☑	gi|73983772|ref|XP_854977.1|	PREDICTED: similar to Ferritin l...	58.9	9e-08
NEW	☑	gi|106895228|ref|ZP_01362330.1|	Ferritin and Dps [Clostridium...	58.5	1e-07
NEW	☑	gi|120528|sp|P18686|FRIL_SHEEP	Ferritin light chain (Ferritin L	58.5	1e-07
NEW	☑	gi|45358722|ref|NP_988279.1|	Ferritin [Methanococcus maripalu...	58.5	2e-07
NEW	☑	gi|288834|emb|CAA47982.1|	ferritin 1 [Vigna unguiculata]	58.2	2e-07
NEW	☑	gi|54113875|gb|AAV29571.1|	NT02FT1049 [synthetic construct]	57.4	3e-07
NEW	☑	gi|56707774|ref|YP_169670.1|	Ferritin-like protein [Francisel...	57.4	3e-07
NEW	☑	gi|90590871|ref|ZP_01246517.1|	Ferritin and Dps [Flavobacteri...	57.0	4e-07
NEW	☑	gi|19553724|ref|NP_601726.1|	ferritin-like protein [Corynebac...	57.0	4e-07
NEW	☑	gi|30023087|ref|NP_834718.1|	Ferritin [Bacillus cereus ATCC 1...	57.0	4e-07
NEW	☑	gi|118170246|gb|ABK71142.1|	ferritin family protein [Mycobacteri	57.0	4e-07
NEW	☑	gi|485051|pir||PQ0614	ferritin 2 - cowpea (fragment)	57.0	4e-07
NEW	☑	gi|84517979|ref|ZP_01005328.1|	ferritin [Prochlorococcus mari...	57.0	4e-07
NEW	☑	gi|89209389|ref|ZP_01187815.1|	Ferritin and Dps [Bacillus wei...	56.6	5e-07
NEW	☑	gi|116074582|ref|ZP_01471843.1|	ferritin [Synechococcus sp. R...	56.6	5e-07
NEW	☑	gi|30265097|ref|NP_847474.1|	ferritin [Bacillus anthracis str...	56.6	6e-07
NEW	☑	gi|53714336|ref|YP_100328.1|	ferritin A [Bacteroides fragilis...	56.6	6e-07
NEW	☑	gi|62263156|gb|AAX78137.1|	unknown protein [synthetic construct]	56.2	6e-07
NEW	☑	gi|18148456|dbj|BAB83264.1|	ferritin-like protein [Clostridium s	56.2	6e-07
NEW	☑	gi|29346517|ref|NP_810020.1|	ferritin A [Bacteroides thetaiot...	55.8	9e-07
NEW	☑	gi|116671801|ref|YP_832734.1|	Ferritin, Dps family protein [A...	55.5	1e-06
NEW	☑	gi|67938798|ref|ZP_00531317.1|	Ferritin and Dps [Chlorobium p...	55.5	1e-06
NEW	☑	gi|39996409|ref|NP_952360.1|	ferritin [Geobacter sulfurreduce...	55.5	1e-06
NEW	☑	gi|92908944|ref|ZP_01277722.1|	Ferritin and Dps [Mycobacteriu...	55.5	1e-06
NEW	☑	gi|42525965|ref|NP_971063.1|	ferritin, putative [Treponema de...	55.5	1e-06
NEW	☑	gi|52140475|ref|YP_086354.1|	ferritin [Bacillus cereus E33L] ...	55.1	1e-06
NEW	☑	gi|47569923|ref|ZP_00240593.1|	SA1709 [Bacillus cereus G9241]...	55.1	2e-06

Run PSI-Blast iteration 2


Run PSI-Blast iteration 3
Hit list size 500
Distance tree of results NEW
Sequences with E-value BETTER than threshold
Related Structures
Sequences producing significant alignments:
Score (Bits)
E Value
gi|47125326|gb|AAH70494.1| FTH1 protein [Homo sapiens] 334 9e-91 U G
gi|114637918|ref|XP 001140124.1| PREDICTED: similar to Ferrit... 332 4e-90 G
gi|76779199|gb|AAI05803.1| FTH1 protein [Homo sapiens] 328 7e-89 U G
gi|83404987|gb|AAI11079.1| Fth1 protein [Rattus norvegicus] 323 2e-87 U G
gi|50927649|gb|AAH78892.1| Fth1 protein [Rattus norvegicus] 323 2e-87 U G
gi|58477732|gb|AAH89817.1| Fth1 protein [Rattus norvegicus] >... 323 3e-87 U G
gi|51859472|gb|AAH81845.1| Fth1 protein [Rattus norvegicus] 322 4e-87 U G
gi|117558589|gb|AAI27508.1| Fth1 protein [Rattus norvegicus] 322 5e-87 G
gi|229918|pdb|1FHA| Chain , Ferritin (H-Chain) Mutant (Lys ... 314 1e-84 S
gi|56682959|ref|NP 002023.2| ferritin, heavy polypeptide 1 [H... 313 2e-84 U G
gi|62900172|sp|Q5R8J7|FRIH PONPY Ferritin heavy chain (Ferrit... 313 2e-84
gi|17367250|sp|Q9XT73|FRIH TRIVU Ferritin heavy chain (Ferrit... 312 6e-84
gi|50978756|ref|NP 001003080.1| ferritin, heavy polypeptide 1... 311 9e-84 U G



	✓	gi|30265097|ref|NP_847474.1|	ferritin [Bacillus anthracis str...	119	8e-26
	✓	gi|76676681|ref|XP_870469.1|	PREDICTED: similar to EGF-like m...	118	1e-25
	✓	gi|52140475|ref|YP_086354.1|	ferritin [Bacillus cereus E33L] ...	118	1e-25
NEW	✓	gi|88711313|ref|ZP_01105401.1|	RsgA [Flavobacteriales bacteri...	118	1e-25
NEW	✓	gi|88804213|ref|ZP_01119733.1|	ferritin 1 [Robiginitalea bifo...	118	1e-25
NEW	✓	gi|47569927|ref|ZP_00240593.1|	SA1709 [Bacillus cereus G9241]...	117	2e-25
	✓	gi|67918558|ref|ZP_00512155.1|	Ferritin and Dps [Chlorobium l...	117	2e-25
NEW	✓	gi|113461064|ref|YP_719132.1|	ferritin like protein 2 [Haemop...	117	3e-25
NEW	✓	gi|57241274|ref|ZP_00369221.1|	ferritin [Campylobacter lari R...	117	3e-25
NEW	✓	gi|15602532|ref|NP_245604.1|	RsgA [Pasteurella multocida subs...	117	3e-25
NEW	✓	gi|89341529|ref|ZP_01193772.1|	Ferritin and Dps [Mycobacteriu...	117	3e-25
NEW	✓	gi|83016346|dbj|BAE53405.1|	ferritin like protein-2 [Actinobacil	116	4e-25
NEW	✓	gi|34540987|ref|NP_905466.1|	ferritin [Porphyromonas gingival...	116	4e-25
	✓	gi|89209389|ref|ZP_01187815.1|	Ferritin and Dps [Bacillus wei...	116	5e-25
NEW	✓	gi|32030887|ref|ZP_00133622.1|	COG1528: Ferritin-like protein [H	116	6e-25
NEW	✓	gi|68250098|ref|YP_249210.1|	ferritin like protein 2 [Haemoph...	116	6e-25
NEW	✓	gi|16273294|ref|NP_439537.1|	ferritin [Haemophilus influenzae...	115	6e-25
	✓	gi|118170246|gb|ABK71142.1|	ferritin family protein [Mycobacteri	115	7e-25
	✓	gi|116671801|ref|YP_832734.1|	Ferritin, Dps family protein [A...	115	7e-25
NEW	✓	gi|42629321|ref|ZP_00154868.1|	COG1528: Ferritin-like protein...	115	1e-24
NEW	✓	gi|113968478|ref|YP_732271.1|	Ferritin, Dps family protein [S...	115	1e-24
NEW	✓	gi|117619373|ref|YP_854579.1|	ferritin [Aeromonas hydrophila ...	115	1e-24
	✓	gi|108873403|gb|EAT37628.1|	ferritin subunit, putative [Aedes ae	114	1e-24
NEW	✓	gi|57242023|ref|ZP_00369963.1|	ferritin VC0078 [Campylobacter...	114	2e-24
NEW	✓	gi|89893918|ref|YP_517405.1|	hypothetical protein DSY1172 [De...	114	2e-24
NEW	✓	gi|53732651|ref|ZP_00154869.2|	COG1528: Ferritin-like protein [H	114	2e-24
NEW	✓	gi|57867357|ref|YP_189000.1|	ferritin family protein [Staphy...	114	2e-24

Run PSI-Blast iteration 3

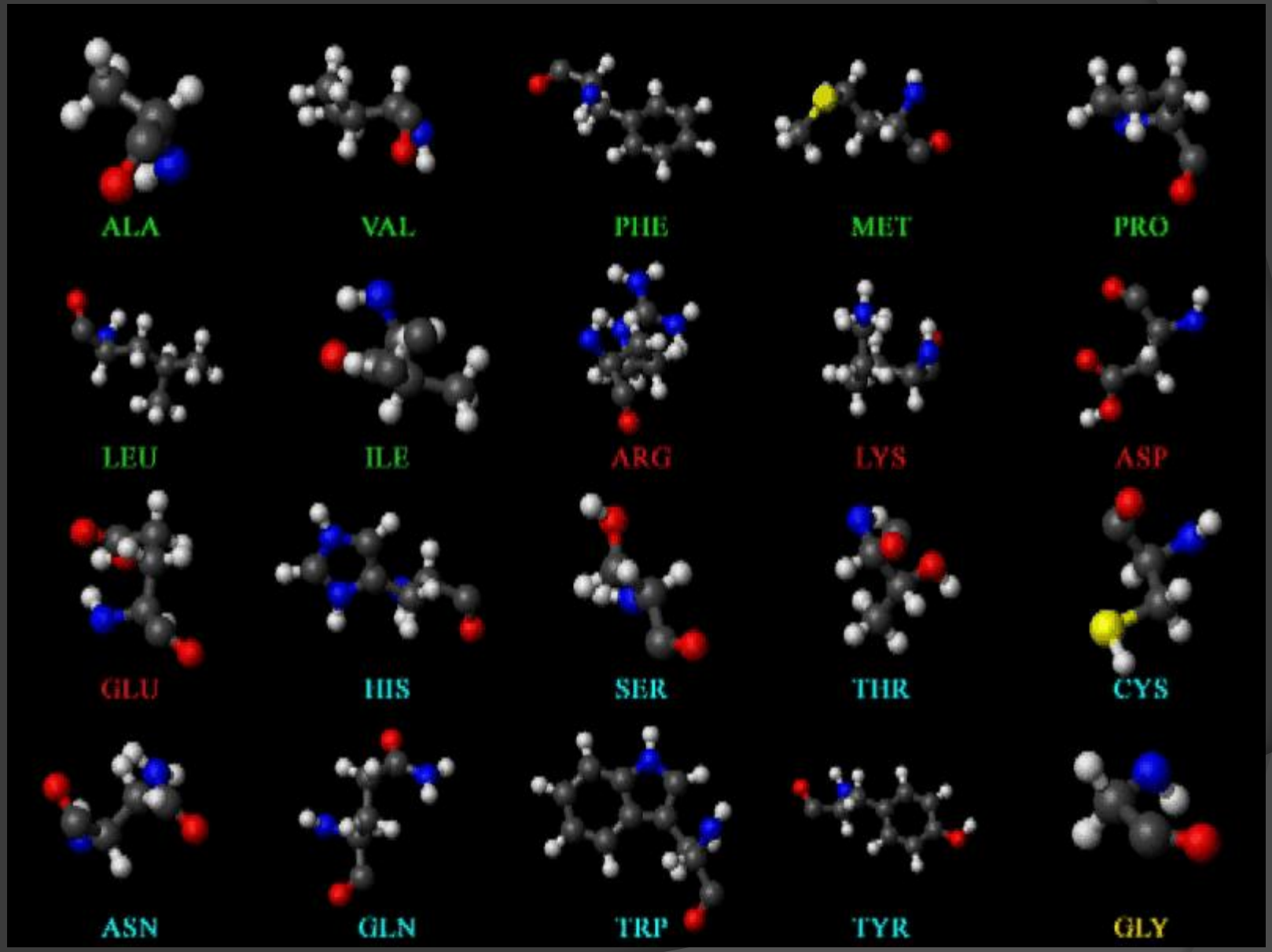
La struttura 3D è responsabile della loro funzione

La pressione evolutiva atta a preservare la funzione ha un effetto importante in contrasto con altri polimeri (formanti globuli amorfi), le proteine hanno un'unica struttura tridimensionale specificata dalla sequenza amminoacidica

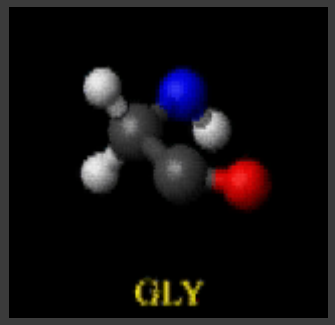
Esempio:

Attività catalitiche (enzimi che espongono siti attivi in un determinato modo)

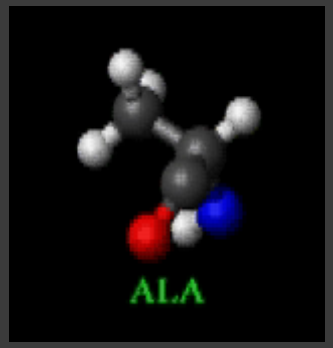
Riconscimento molecolare (immunoglubilne)


ALA
VAL
PHE
MET
PRO
LEU
ILE
ARG
LYS
ASP
GLU
HIS
SER
THR
CYS
ASN
GLN
TRP
TYR
GLY

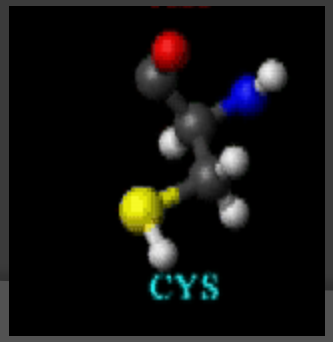
Ogni aa ha una sua particolare funzione sia a livello strutturale che di funzione



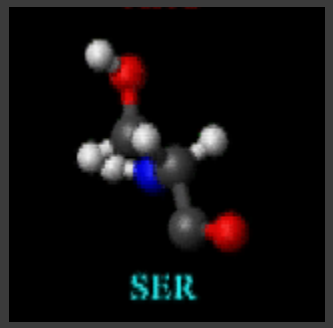
Molto flessibile (ϕ e ψ) si trova soprattutto in corrispondenza di turn, rara in eliche e sheet per l'assenza di legami H



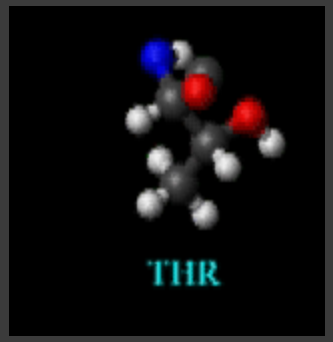
Idrofobico ma allo stesso tempo può trovarsi esposto al solvente pertanto tipico nelle zone di interazione tra proteine



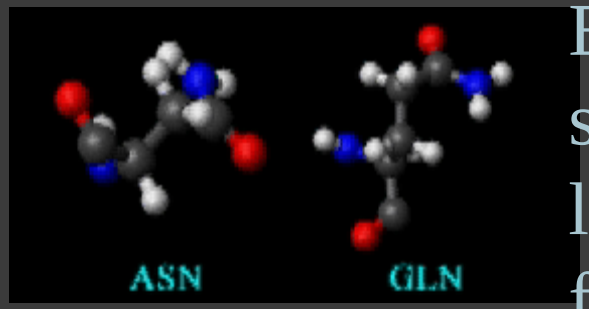
Ponti a disolfuro, coordinante metalli, attività redox (albumina)



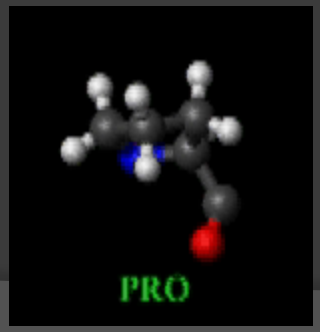
Dentro e fuori le proteine può essere associato a deboli turns per la flessibiltà e la capacità di formare legami H con il bb, può essere fosforilata…cascata di traduzione del segnale



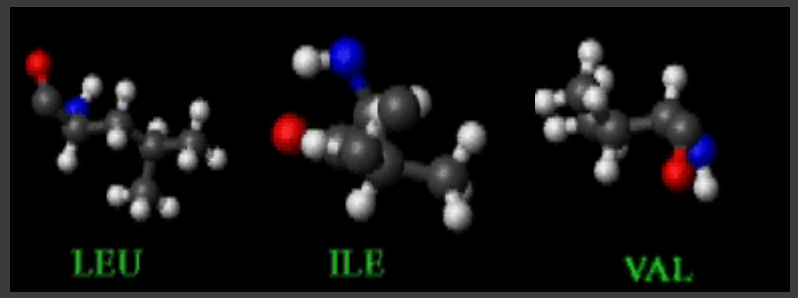
Proprietà simili alla Ser ma essendo ramificato è meno flessibile difficile da posizionarsi nelle eliche comune negli sheet.



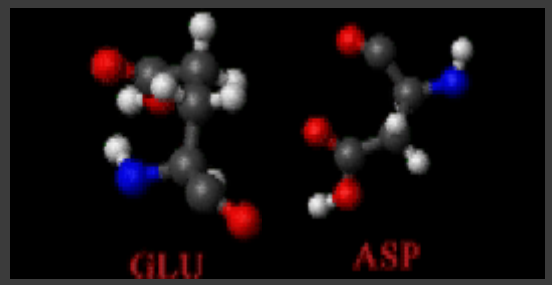
Entrambi polari si trovano di solito sulla superficie e siti attivi, Asn può formare legami H con il bb e indurre la formazione di eliche



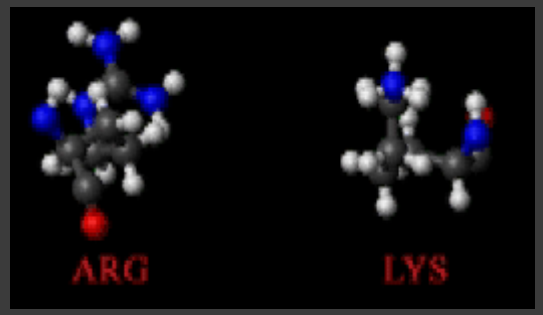
E’ unico perché imminoacido, ciò restringe la sua flessibiltà impedendogli di formare alcuni legami H induce un ripiegamento, nelle eliche…



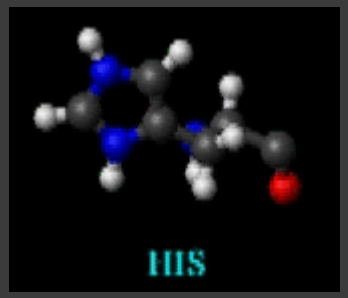
Idrofobici raramente in siti attivi tipici degli sheet.



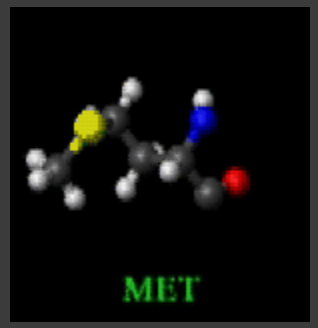
Carichi – di solito esposti se interni formano ponti salini con aa carichi +, siti attivi e coordinanti metalli (Zn)



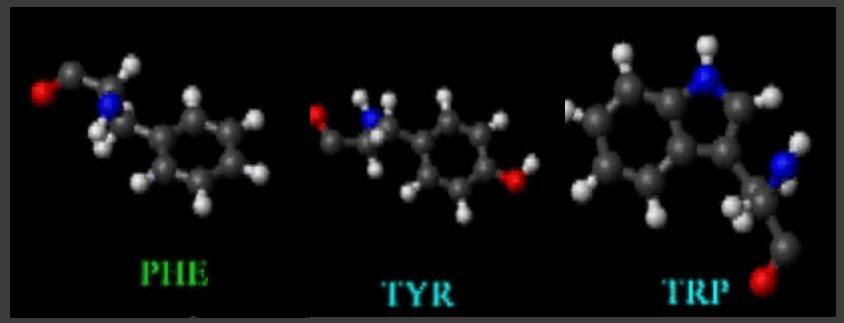
Carichi + hanno la prima parte della catena idrofobica, possono formare ponti salini e interagire con fosfati (ATP)



A pH fisiologico può agire come base o acido, residuo ideale per centri funzionali, può inoltre coordinare metalli (His Tag)



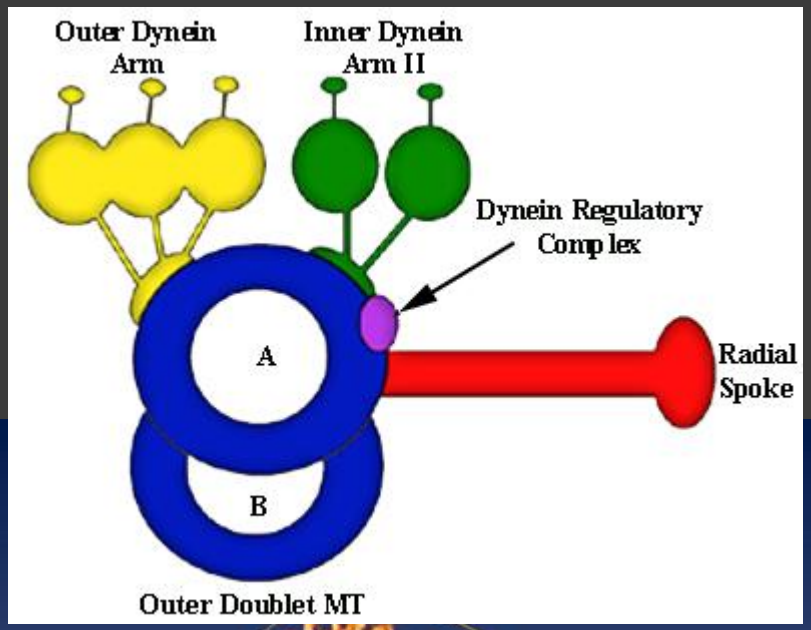
Lunga sc idrofobica e flessibile, dentro le proteine, a un atomo di S legato però ad un gruppo metile che lo rende meno reattivo.

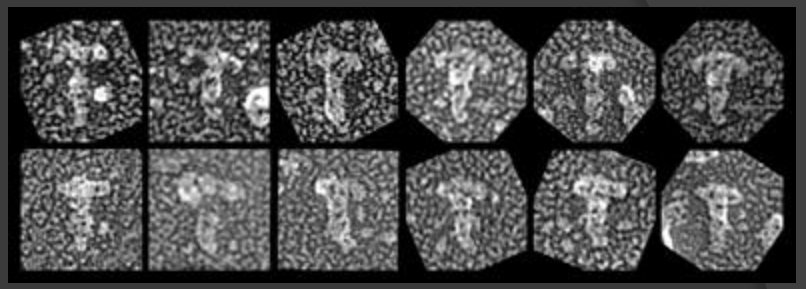


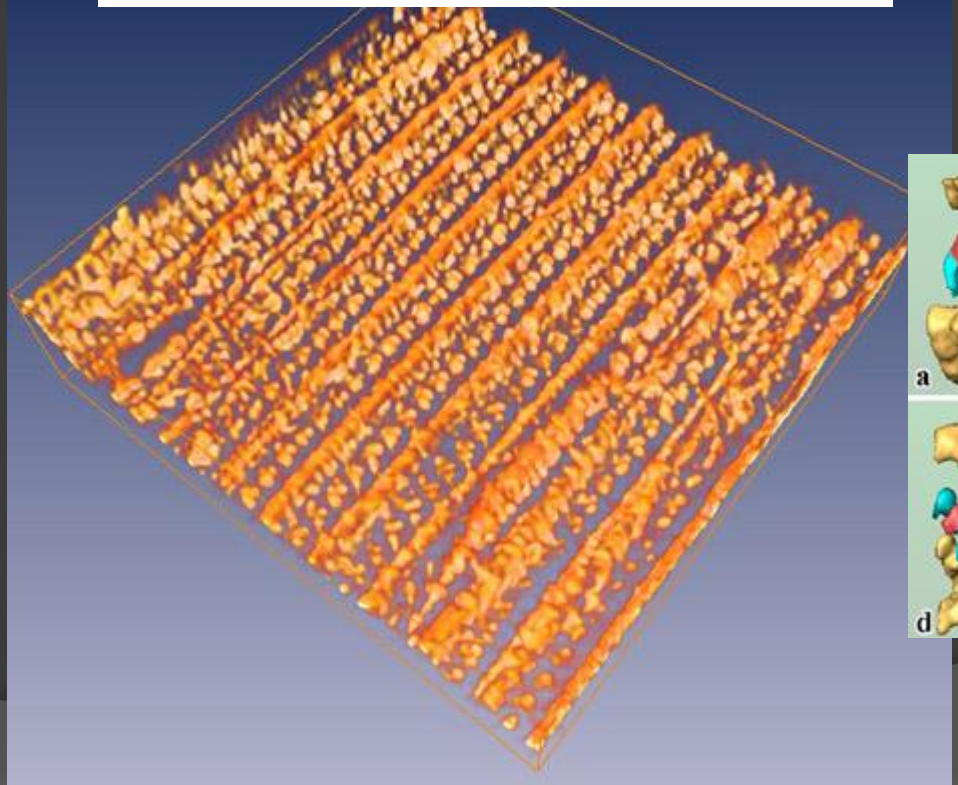
Aromatici.

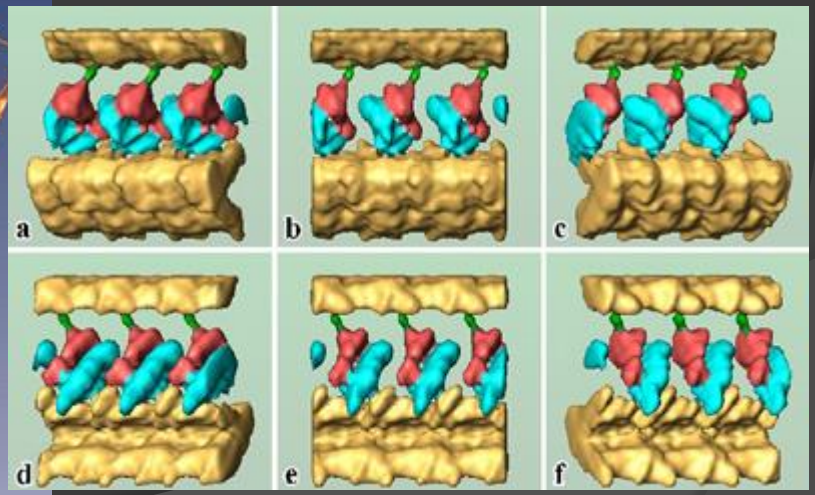
La struttura quaternaria di grandi molecole proteiche può essere studiata in 3D a pertire da proiezioni ottenute al TEM lungo direzioni casuali.

Transmission Electron Microscopy -> La Microscopia Elettronica a Trasmissione (TEM) permette di ottenere, da un campione sufficientemente assottigliato (< 0.1 micron), immagini ad alta risoluzione (< 10 Å) prodotte da elettroni ad alta energia (100 KeV) trasmessi su uno schermo fluorescente o su una pellicola fotografica.


Outer Dynein Arm
Inner Dynein Arm II
Dynein Regulatory Complex
A
B
Radial Spoke
Outer Doublet MT






a
b
c
d
e
f

Il problema del protein folding…
$\Delta G=\Delta H-T\Delta S$ differenza tra lo stato foldato e non

Le informazioni contenute nella sequenza proteica sono sufficienti a specificare la struttura

La Structural Genomics Initiative si ripromette di determinare tra 10000 e 20000 nuove strutture proteiche nei prossimi anni e la maggior parte di esse saranno MODELLATE

La sfida ovviamente è quella di determinarne la funzione dopo la struttura.

Proteine omologhe hanno la medesima struttura
Occasionalmente proteine che non sembrano avere un evoluzione comune hanno la stessa topologia.

Potremmo pensare di ottenere informazioni sulla funzione dall'analisi per esempio di proteine a struttura similare, ma ci sono almeno tre difficoltà:

Proteine omologhe potrebbero essere originate da duplicazione genica con differente evoluzione ed avere acquisito differenti funzioni

Alcuni ripiegamenti sono adottati da proteine che mostrano differenti funzioni

La proteina considerata potrebbe avere un ripiegamento nuovo non ancora osservato

Cosa possiamo imparare però dall'analisi della struttura?

Quali residui sono esposti e quali no
La struttura quaternaria
La struttura ottenuta sarà sicuramente molto simile a quella biologicamente attiva.

Per esempio helix-turn-helix motifs > lega il DNA
Due eliche attorcigliate per circa otto giri con Leu presenti ogni sette residui indicano un sito di dimerizzazione di molte proteine leganti il DNA
Un motivo in cui lo Zn è legato a due cys e due his separate da dodici residui zink-finger

Nel caso di assenza di motivi funzionali possiamo comunque analizzare la presenza di crepe sulla superficie ed evidenziare possibili siti catalitici, non è però sempre possibile, a meno che non si conoscano altre proteine della stessa famiglia evolutiva.

Recentemente si è visto come molte proteine hanno più di una attività

Altre hanno mostrato di ripiegarsi correttamente solo quando si legano ad uno specifico ligando.

153320+1062164 sequenze proteiche conosciute SwissProt +Trembl

26059 strutture conosciute

Tecniche lunghe per determinare la struttura delle proteine (NMR, Cristallografia, TEM)

Incapacità di prevedere il folding di una proteina

Questo ha portato alla necessità di sviluppare una nuova metodologia.

Predizione struttura proteine:

Per farlo bisogna dividere il procedimento in step successivi

Struttura secondaria è predetta a partire dalla sequenza primaria

Gli elementi di struttura secondaria si riarrangiano per formare la struttura terziaria

Struttura secondaria
Perché è così complicato?

Per una sequenza di 100 aa, se assumiamo solo 2 possibili conformazioni per ogni residuo, ci sono $2^{100} \sim 10^{30}$ per l'intera catena

Dal momento che la sequenza determina la struttura bisogna determinare la struttura secondaria a partire dalla sequenza

Naturalmente la predizione non è così semplice

ci sono casi in cui:

Sequenze simili danno differenti strutture

Mutazioni puntiformi possono alterare la struttura

Differenti sequenze danno strutture simili

Globin fold

Predizione Struttura Secondaria

I metodi usati sono tre e si basano sulle informazione raccolte dalle proteine la cui struttura terziaria è già risolta.

-Statistico di Chou e Fasman: i 20 aa mostrano preferenze significative per particolari strutture secondarie
(A,R,Q,E,M,L,K eliche) (C,I,F,T,W,Y,V foglietti)
GOR attendibilità del 56%

-Stereochimico di Lim: tiene conto delle proprietà idrofobiche, idrofiliche ed elettrostatiche considerando il loro ruolo nel folding (alternanza di idrofilici e idrofobici, foglietti) utile per predire eliche anfipatiche e transmembrana. SOSUI, TMPRED, ecc.

-Neural Network: tiene conto di entrambe le precedenti e del processo evolutivo a partire dall'allineamento multiplo. PHD 70%

Chou-Fasman / GOR Method

Prediction: predicted sequence is scanned

- α helix prediction: when 4/6 a.a. have a probability > 1.03 to be α helix.
- β sheet prediction: when 3/5 a.a. have a probability > 1 to be β sheet.
- prediction is elongated until prediction values of 4 a.a. are < 1.

PSORT.org provides links to the PSORT family of programs for subcellular localization prediction as well as other datasets and resources relevant to localization prediction. The page is currently hosted by the Brinkman Laboratory at Simon Fraser University, and our goal is to provide an open-source resource centre for researchers interested in subcellular localization prediction.

Please choose from the following PSORT programs for localization prediction:

- PSORTb v.2.0 (Gardy et al, 2004) (v.1.0: Gardy et al, 2003) for **bacterial sequences**
- WoLF PSORT (Horton et al., to be published) is a **recently updated version of PSORT II** for the prediction of **eukaryotic sequences**
- PSORT II (Nakai and Horton, 1997) for **eukaryotic sequences**
- PSORT (Nakai and Kanehisa, 1991) for **plant sequences**
- iPSORT (Bannai et al, 2002) for **classification of eukaryotic N-terminal sorting signals**

See also PSORTdb, our new database of bacterial protein subcellular localizations.

PSORTb and PSORTdb are maintained by the Brinkman Laboratory, Simon Fraser University, British Columbia, Canada. PSORT and PSORT II are maintained by Kenta Nakai, at the Human Genome Center, Institute for Medical Science, University of Tokyo, Japan. iPSORT is maintained by Hideo Bannai at the Human Genome Center.

PSORT-B Results

```
SeqID: ecoli
  Analysis Report:
    SCL-BLASTe-       Periplasmic, CytoplasmicMembrane[matched 100% 129663: Cytoplasmic membrane associated periplasmic protein]
  Localization Scores:
    Cytoplasmic             0.00
    CytoplasmicMembrane     10.00
    Periplasmic             10.00
    OuterMembrane           0.00
    Extracellular           0.00
  Final Prediction:
    CytoplasmicMembrane (This protein may have multiple localization sites.) 10.00
--------------------------------------------------------------------------------
```

TargetP 1.1 Server

Predice la localizzazione subcellulare di proteine eucariotiche .
Testato per la prima volta su Arabidopsis Thaliana (per le piante) e Homo Sapiens (per le non piante).

La localizzazione è basato sulla presenza delle presequenze N-terminali, cioè:

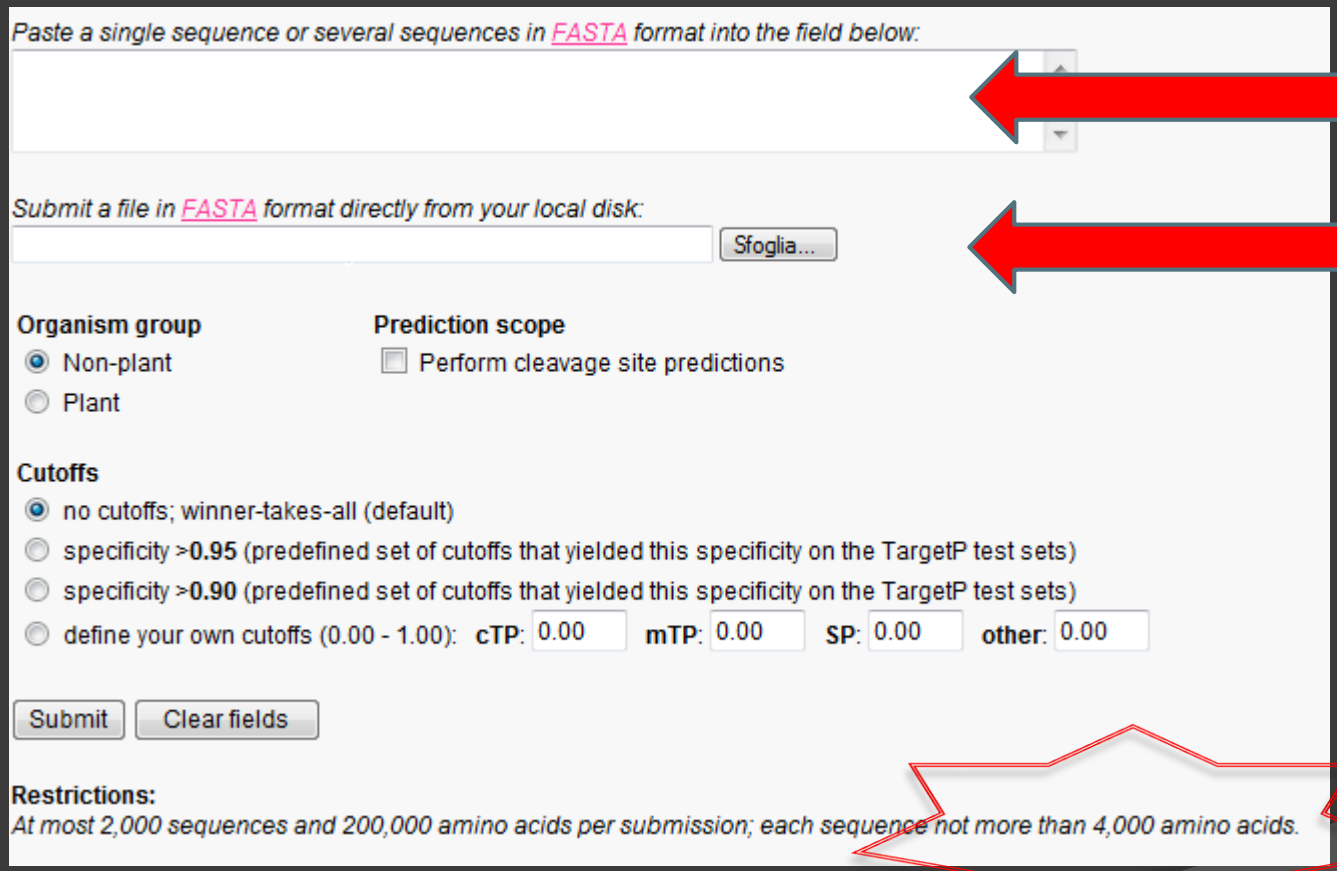
- ***peptide targeting di cloroplasti (cTP),*** per tale predizione è utilizzato **ChloroP** che predice la presenza di sequenze segnale e la potenziale localizzazione del sito di taglio

- ***peptide targeting di mitocondri (mTP)***

- ***peptide segnale della via secretoria (SP)***, per tale predizione si usa **SignalP** che come **ChloroP** predice la presenza di sequenze segnale e la potenziale localizzazione del sito di taglio, ciò lo esegue in sequenze aminoacidiche di organismi diversi.

Come si usa TargetP???

1-Inserire i dati



Codice ad una lettera

Percorso di directory

Paste a single sequence or several sequences in FASTA format into the field below:

Submit a file in FASTA format directly from your local disk:

Sfoglia...

Organism group

- (•) Non-plant
- ( ) Plant

Prediction scope

- [ ] Perform cleavage site predictions

Cutoffs

- (•) no cutoffs; winner-takes-all (default)
- ( ) specificity >**0.95** (predefined set of cutoffs that yielded this specificity on the TargetP test sets)
- ( ) specificity >**0.90** (predefined set of cutoffs that yielded this specificity on the TargetP test sets)
- ( ) define your own cutoffs (0.00 - 1.00): **cTP**: 0.00 **mTP**: 0.00 **SP**: 0.00 **other**: 0.00

Submit | Clear fields

Restrictions:
At most 2,000 sequences and 200,000 amino acids per submission; each sequence not more than 4,000 amino acids.

Demanded specificity	Category	Cutoff	
		Plant	Non-plant
0.95	cTP	0.73	-
	mTP	0.86	0.78
	SP	0.43	0.00
	other	0.84	0.73
0.90	cTP	0.62	-
	mTP	0.76	0.65
	SP	0.00	0.00
	other	0.53	0.52

Si può definire anche il range di cutoffs per cTP, mTP e SP che va da 0,00 a 1,00 (range valori di output)

Risultati

Nome max 20 caratteri

Lunghezza sequenza

Predizione di localizzazione basata sugli score

C,M,S,-,*

Reliability class

```
### targetp v1.1 prediction results ###################################
Number of query sequences:  12
Cleavage site predictions included.
Using PLANT networks

Name                  Len     cTP    mTP     SP  other  Loc  RC  TPlen
----------------------------------------------------------------------
P11043_has_a_very_ve  516   0.873  0.012  0.004  0.320   C    3     65
P07505                266   0.330  0.047  0.004  0.444   _    5      -
P12360                246   0.580  0.119  0.210  0.089   C    4     42
P12352                 97   0.397  0.555  0.014  0.150   M    5     40
Q01289                399   0.733  0.017  0.031  0.462   C    4     62
P08817                129   0.844  0.092  0.089  0.015   C    2     47
P07263                546   0.400  0.380  0.075  0.020   C    5     41
P07597                117   0.005  0.095  0.967  0.006   S    1     26
P48786               1088   0.199  0.070  0.067  0.822   _    2      -
Q01238                102   0.420  0.277  0.033  0.164   C    5     41
P35334                342   0.055  0.010  0.968  0.041   S    1     29
P13086                333   0.053  0.905  0.045  0.034   M    1     21
----------------------------------------------------------------------
cutoff                      0.000  0.000  0.000  0.000
```

Predizione lunghezza presequenza, mostarta solo quando a TargetP si chiede la localizzazione del sito di taglio

Score finale su cui è basata la predizione finale. La localizzazione con un alto score è quella più in accordo con TargetP

Esempi

>P16096; 46 FRUCTOSE-BISPHOSPHATE ALDOLASE, CHLOROPLAST PRECURSOR

```
### targetp v1.1 prediction results ###################################
Number of query sequences:  1
Cleavage site predictions not included.
Using PLANT networks.

Name                    Len      cTP    mTP     SP   other  Loc  RC
----------------------------------------------------------------------
394                       0    0.090  0.112  0.951  0.000
----------------------------------------------------------------------
cutoff                         0.000  0.000  0.000  0.000
```

>P00508; 22 ASPARTATE AMINOTRANSFERASE, MITOCHONDRIAL PRECURSOR

```
### targetp v1.1 prediction results ###################################
Number of query sequences:  1
Cleavage site predictions not included.
Using NON-PLANT networks.

Name                    Len             mTP     SP   other  Loc  RC
----------------------------------------------------------------------
423                       0           0.031  0.978  0.000
----------------------------------------------------------------------
cutoff                                0.000  0.000  0.000
```

Predizioni di struttura secondaria

CHOU & FASMAN

Residue	helix	ß-sheet	turns
Glu	1.59	0.52	1.01
Ala	1.41	0.72	0.82
Leu	1.34	1.22	0.57
Met	1.30	1.14	0.52
Gln	1.27	0.98	0.84
Lys	1.23	0.69	1.07
Arg	1.21	0.84	0.90
His	1.01	0.80	0.81
Val	0.90	1.87	0.41
Ile	1.09	1.67	0.47
Tyr	0.74	1.45	0.76
Cys	0.66	1.40	0.54
Trp	1.02	1.35	0.65
Phe	1.16	1.33	0.59
Thr	0.76	1.17	0.90
Gly	0.43	0.58	1.77
Asn	0.76	0.48	1.34
Pro	0.34	0.31	1.32
Ser	0.57	0.96	1.22
Asp	0.99	0.39	1.24

Pôle BioInformatique Lyonnais
Network Protein Sequence Analysis

NPS@ is the IBCP contribution to PBIL in Lyon, France

Wednesday, December 28th 2005: NPS@ server will not be available from January 9th to 13th. We apologize for any inconveniences.

Monday, September 26th 2005 : fixed secondary structure prediction insertion in Multalin. (see news)
When sending automatic requests on NPS@, please use HTTP POST method not GET.

GOR IV SECONDARY STRUCTURE PREDICTION METHOD

[Abstract] [NPS@ help] [Original server]

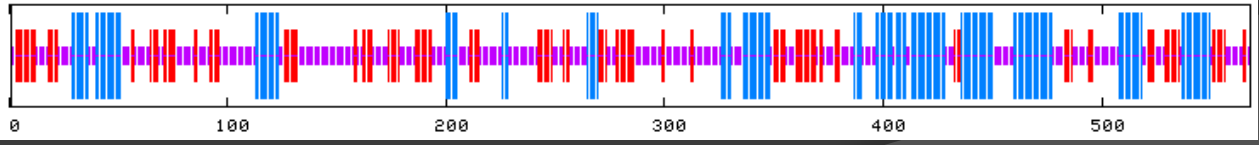
Sequence name (optional) :

Paste a protein sequence below : help

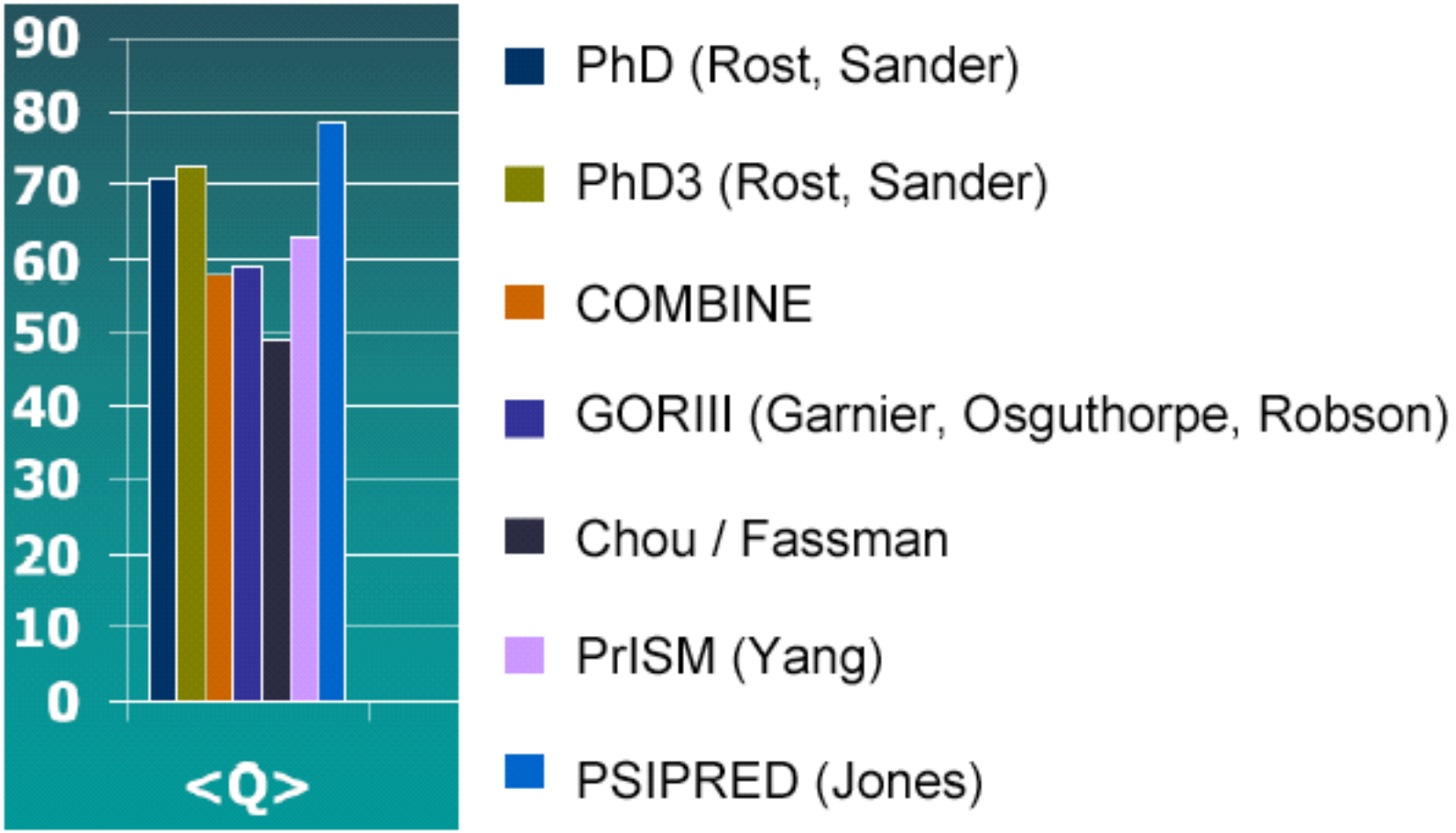
```
NGVKPLILRDCSVAGWLLGNPMCDEFINVPEWSYIVEKASPANDLCYPGNFNDYEELKHL
LSRINHFEKIQIIPKSSWSNHDASSGVSSACPYLGRSSFFRNVVWLIKKNSAYPTIKRSY
NNTNQEDLLVLWGVHHPNDAAEQTKLYQNPTTYISVGTSTLNQRLVPEIATRPKVNGQSG
RMEFFWTILKPNDAINFESNGNFIAPEYAYKIVKKGDSTIMKSELEYGNCNTKCQTPMGA
INSSMPFHNIHPLTIGECPKYVKSNRLVLATGLRNTPQRERRRKKRGLFGAIAGFIEGGW
QGMVDGWYGYHHSNEQGSCYSADKESTQKAIDGVTNKVNSIINKMNTQFEAVGREFNNLE
RRIENLNKKMEDGFLDVWTYNAELLVLMENERTLDFHDSNVKNLYDKVRLQLRDNAKELG
NGCFEFYHKCDNECMESVKNGTYDYPQYSEEARLNREEISGVKLESMGTYQILSIYSTVA
SSLALAIMVAGLSLWMCSNGSLQCRICI
```

Output width : 70

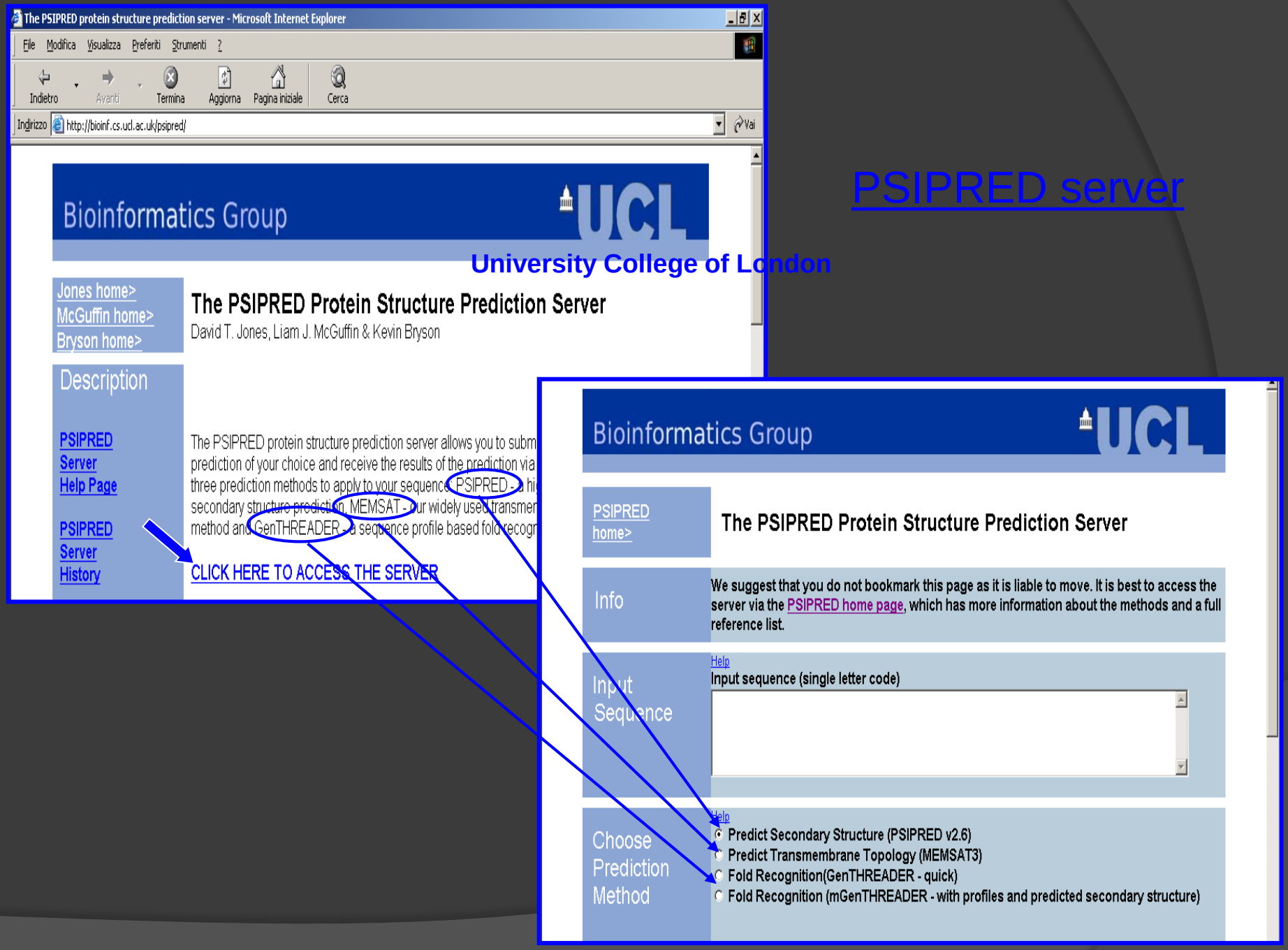
```
         10         20         30         40         50         60         70
          |          |          |          |          |          |          |
MERTVLLLATVSLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQDILERTHNGKLCDLNGVKPLILRD
ccceeeeeeeeeccccceeeeeccccchhhhhhhhcccchhhhhhhhhhhhhccccceeccccccceeeec
CSVAGWLLGNPMCDEFINVPEWSYIVEKASPANDLCYPGNFNDYEELKHLLSRINHFEKIQIIPKSSWSN
ceeeeeeccccccccceecccccceeeeeccccccccccccccchhhhhhhhhhhccceeeeeeeccccccc
HDASSGVSSACPYLGRSSFFRNVVWLIKKNSAYPTIKRSYNNTNQEDLLVLWGVHHPNDAAEQTKLYQNP
ccccccccccccccccccceeccceeeeeccccccceeeeeccccccceeeeeeeeccccccchhhhhhcccc
TTYISVGTSTLNQRLVPEIATRPKVNGQSGRMEFFWTILKPNDAINFESNGNFIAPEYAYKIVKKGDSTI
ceeeeeccccccccccchhhccccccccccccceeeeeeeccccceeecccccccchhhhheeeeccccee
MKSELEYGNCNTKCQTPMGAINSSMPFHNIHPLTIGECPKYVKSNRLVLATGLRNTPQRERRRKKRGLFG
eeeeeeecccccccccccceecccccccccceecccccccccccchhhhhccccchhhhhhhhhhhhhc
AIAGFIEGGWQGMVDGWYGYHHSNEQGSCYSADKESTQKAIDGVTNKVNSIINKMNTQFEAVGREFNNLE
eeeeeecccceeeeeeeeeeeecccccceeecccccchhhhcccccchhhhhhhhchhhhhcchhhhhhh
RRIENLNKKMEDGFLDVWTYNAELLVLMENERTLDFHDSNVKNLYDKVRLQLRDNAKELGNGCFEFYHKC
hhhhhhhhhccccceeehhhhhhhhhhhhhhhcccccccccchhhhhhhhhhhhhhhhhhhccccceeeeccc
DNECMESVKNGTYDYPQYSEEARLNREEISGVKLESMGTYQILSIYSTVASSLALAIMVAGLSLWMCSNG
cccceeeccccccccccchhhhhhhhhhhcceeeeccccceeeeeeecchhhhhhhhhhhhhceeeeeeccc
SLQCRICI
ccccceec

Sequence length :   568

GOR4 :
   Alpha helix      (Hh) :   153 is  26.94%
   3_10  helix      (Gg) :     0 is   0.00%
   Pi helix         (Ii) :     0 is   0.00%
   Beta bridge      (Bb) :     0 is   0.00%
   Extended strand  (Ee) :   143 is  25.18%
   Beta turn        (Tt) :     0 is   0.00%
   Bend region      (Ss) :     0 is   0.00%
   Random coil      (Cc) :   272 is  47.89%
   Ambigous states  (?)  :     0 is   0.00%
   Other states          :     0 is   0.00%
```



Algoritmi predizione struttura secondaria



PSIPRED server



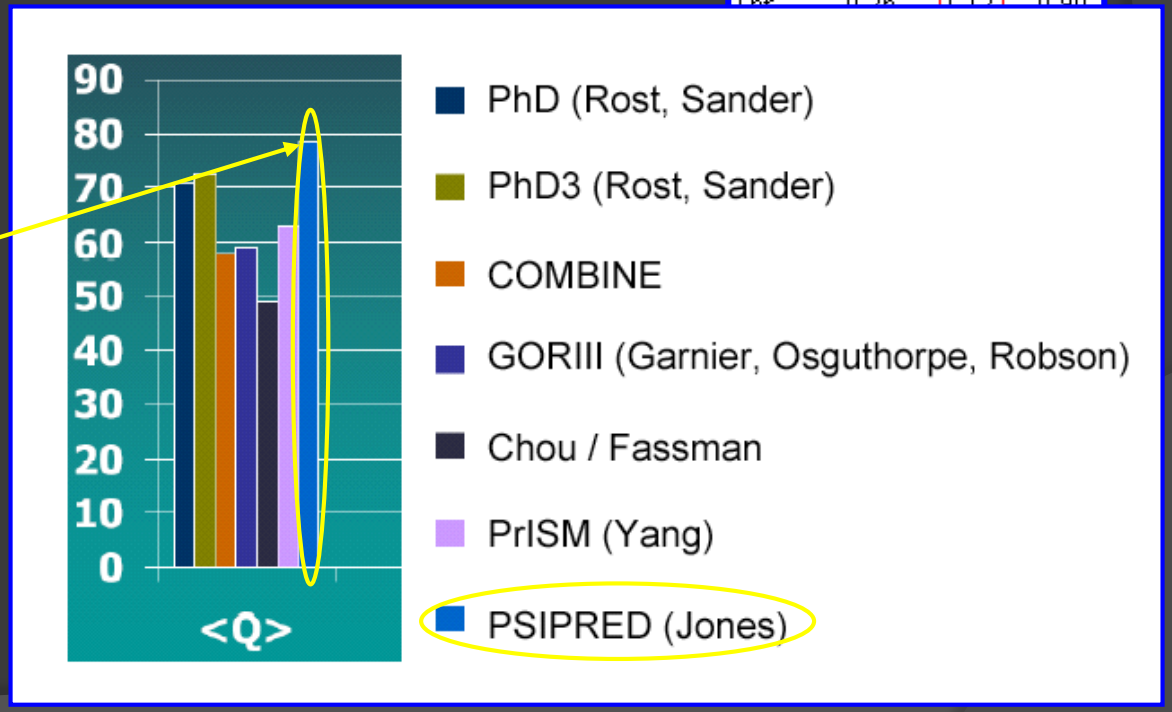
Predizione struttura secondaria

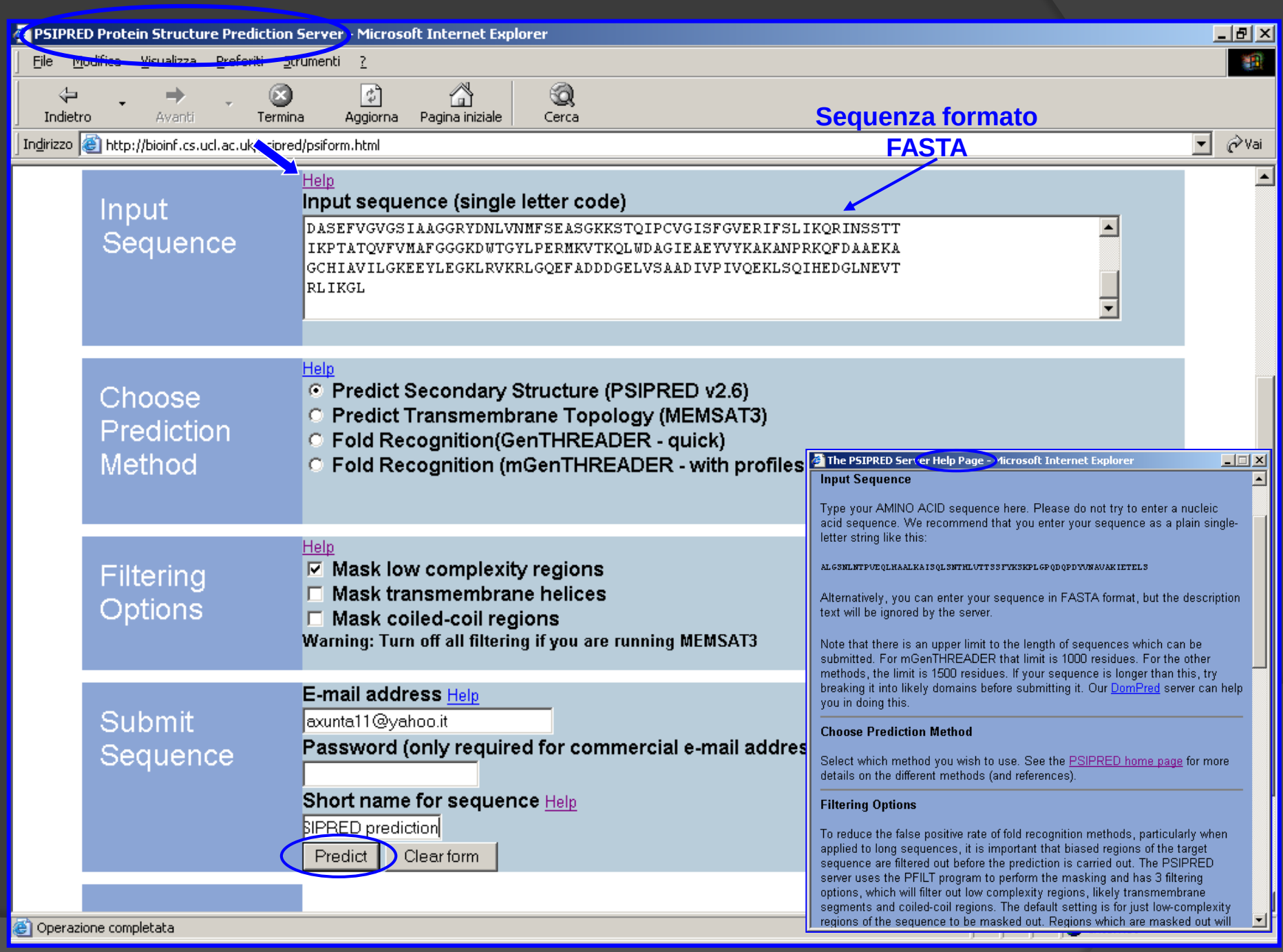
Il metodo PSIPRED tiene conto di:

- metodo statistico di Chou e Fasman;
- proprietà chimico-fisiche degli aa;
- allineamento multiplo effettuato da PSI-BLAST.

Residue	helix	ß-sheet	turns
Glu	1.59	0.52	1.01
Ala	1.41	0.72	0.82
Leu	1.34	1.22	0.57
Met	1.30	1.14	0.52
Gln	1.27	0.98	0.84
Lys	1.23	0.69	1.07
Arg	1.21	0.84	0.90
His	1.01	0.80	0.81
Val	0.90	1.87	0.41
Ile	1.09	1.67	0.47
Tyr	0.74	1.45	0.76
Cys	0.66	1.40	0.54
Trp	1.02	1.35	0.65
Phe	1.16	1.33	0.59

Attendibilità dei risultati molto elevata




PSIPRED Protein Structure Prediction Server - Microsoft Internet Explorer
File Modifica Visualizza Preferiti Strumenti ?
Indietro
Avanti
Termina
Aggiorna
Pagina iniziale
Cerca
Indirizzo http://bioinf.cs.ucl.ac.uk/psipred/psiform.html
Vai
Sequenza formato FASTA
Input Sequence
Help
Input sequence (single letter code)
DASEFVGVGSIAAGGRYDNLVNMFSEASGKKSTQIPCVGISFGVERIFSLIKQRINSSTT
IKPTATQVFVMAFGGGKDWTGYLPERMKVTKQLWDAGIEAEYVYKAKANPRKQFDAAEKA
GCHIAVILGKEEYLEGKLRVKRLGQEFADDDGELVSAADIVPIVQEKLSQIHEDGLNEVT
RLIKGL
Choose Prediction Method
Help
Predict Secondary Structure (PSIPRED v2.6)
Predict Transmembrane Topology (MEMSAT3)
Fold Recognition(GenTHREADER - quick)
Fold Recognition (mGenTHREADER - with profiles
Filtering Options
Help
Mask low complexity regions
Mask transmembrane helices
Mask coiled-coil regions
Warning: Turn off all filtering if you are running MEMSAT3
Submit Sequence
E-mail address Help
axunta11@yahoo.it
Password (only required for commercial e-mail addres
Short name for sequence Help
SIPRED prediction
Predict
Clear form
Operazione completata
The PSIPRED Server Help Page - Microsoft Internet Explorer
Input Sequence
Type your AMINO ACID sequence here. Please do not try to enter a nucleic acid sequence. We recommend that you enter your sequence as a plain single-letter string like this:
ALGSNLNTPVEQLHAALKAISQLSNTHLVTTSSFYKSKPLGPQDQPDYVNAVAKIETELS
Alternatively, you can enter your sequence in FASTA format, but the description text will be ignored by the server.
Note that there is an upper limit to the length of sequences which can be submitted. For mGenTHREADER that limit is 1000 residues. For the other methods, the limit is 1500 residues. If your sequence is longer than this, try breaking it into likely domains before submitting it. Our DomPred server can help you in doing this.
Choose Prediction Method
Select which method you wish to use. See the PSIPRED home page for more details on the different methods (and references).
Filtering Options
To reduce the false positive rate of fold recognition methods, particularly when applied to long sequences, it is important that biased regions of the target sequence are filtered out before the prediction is carried out. The PSIPRED server uses the PFILT program to perform the masking and has 3 filtering options, which will filter out low complexity regions, likely transmembrane segments and coiled-coil regions. The default setting is for just low-complexity regions of the sequence to be masked out. Regions which are masked out will

Risultati predizione (formato di testo)

PSIPRED PREDICTION RESULTS

Key

Conf: Confidence (0=low, 9=high)
Pred: Predicted secondary structure (H=helix, E=strand, C=coil)
AA: Target sequence

PSIPRED HFORMAT (PSIPRED V2.6 by David Jones)

Conf: 954335641333114877027788842568999999985568789750334412689997
Pred:
CCCCCCCCCEEEHHHHHHHHCCCCCCCHHHHHHHHHHHHHCCCCCCCCHHHHHHCCCCC
CC
AA:
MLSRSLNKVVTSIKSSSIIRMSSATAAATSAPTANAANALKASKAPKKGKLQVSLKTPKG
10 20 30 40 50 60

Conf: 421599999999999999999984987801655450877444456452106999958 9
Pred:
CCCCCHHHHHHHHHHHHHHHHHHHHCCCEEEBCCCCCCHHHHCCCCCCCCCHHEEEEEE
CC
AA:
TKDWADSDMVIREAIFSTLSGLFKKHGGVTIDTPVFELREILAGKYGEDSKLIYNLEDQG
70 80 90 100 110 120

Conf: 9889758856299999999579975016987456844787545886351125177996 48
Pred:
CCEEEECCCCCHHHHHHHHHCCCCCEEEEEBCCEEBCCCCCCCCCCCCCCEEECEEEBCC
AA:
GELCSLRYDLTVPFARYVAMNNIQSIKRYHIAKVYRRDQPAMTKGRMREFYQCDFDVAGT
130 140 150 160 170 180

Conf: 9965879999999999999779987599977856689999984999999999999998 6
Pred:
CCCHHHHHHHHHHHHHHHHHCCCCCEEEEBCCHHHHHHHHHHHCCCCHHHHHHHHHHH
HHH
AA: FESMVPDSECLSILVEGLTSLGIKDFKIKLNHRKILDGIFQIAGVKDEDVRKISSAVDKL
190 200 210 220 230 240

Conf: 4020999999864134878899999988998629878999999846443456658999 99
Pred:
CCCHHHHHHHHHHHHCCCCCHHHHHHHHHHHHHHCCCHHHHHHHHHHHCCCCCCCHHHH

250 260 270 280 290 300

Conf: 9999999999739982799864101587245865999995688886654200025667 55
Pred:
HHHHHHHHHHHHCCCCEEEEEBCCCCCCCCCCCCCCEEEEEEEBCCCCCCCCCCCCCCCCCCCC
AA:
DIATLMKYTEAFDIDSFISFDLSLARGLDYYTGLIYEVVTSASAPPENASELKKKAKSAE
310 320 330 340 350 360

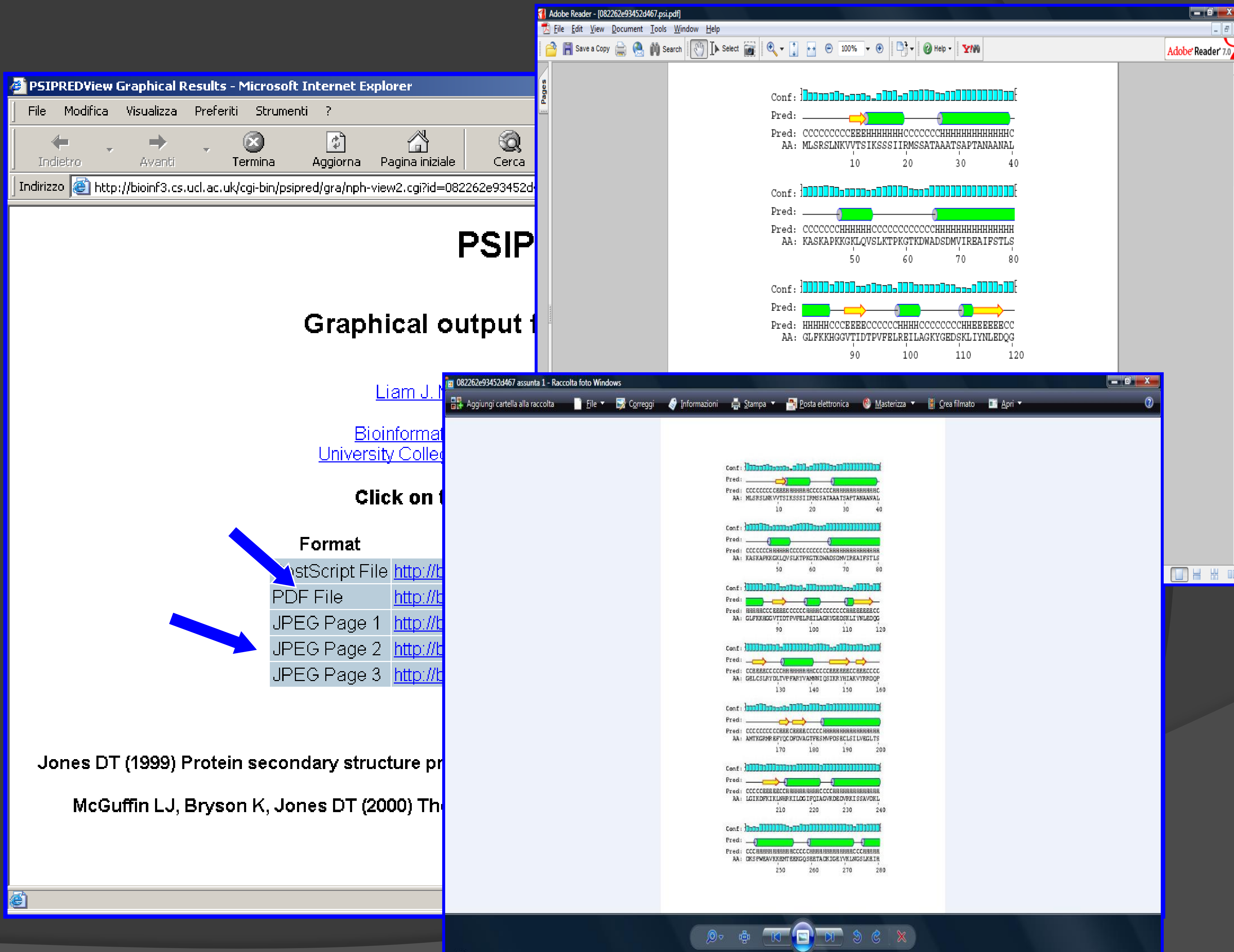
Conf: 4556665525874745778999857000257997774799961999999999753157 76
Pred:
CCCCCCCCCEEEEBCCCHHHHHHHHHCCCCCCCCCCCCEEEEEECHHHHHHHHHHHHHCCCC
C
AA: DASEFVGVGSIAAGGRYDNLVNMFSEASGKKSTQIPCVGISFGVERIFSLIKQRINSSTT
370 380 390 400 410 420

Conf: 6788762999976705667778999999999997798099996888799999999998 7
Pred:
CCCCCCEEEEEEBCCHHHHHHHHHHHHHHHHHHHHHHCCCCEEEEEBCCCCCHHHHHHHHH
HC
AA:
IKPTATQVFVMAFGGGKDWTGYLPERMKVTKQLWDAGIEAEYVYKAKANPRKQFDAAE
KA
430 440 450 460 470 480

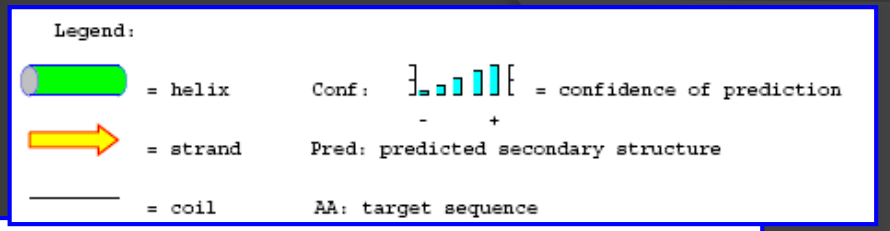
Conf: 9899999876477679699998976456686258759999999999999887235789 9
Pred:
CCCEEEEEBCCCHHHCCEEEEEECCCCCCCCCCEEEEBCHHHHHHHHHHHHHHHHHHHHHHH
H
AA:
GCHIAVILGKEEYLEGKLRVKRLGQEFADDDGELVSAADIVPIVQEKLSQIHEDGLNEVT
490 500 510 520 530 540

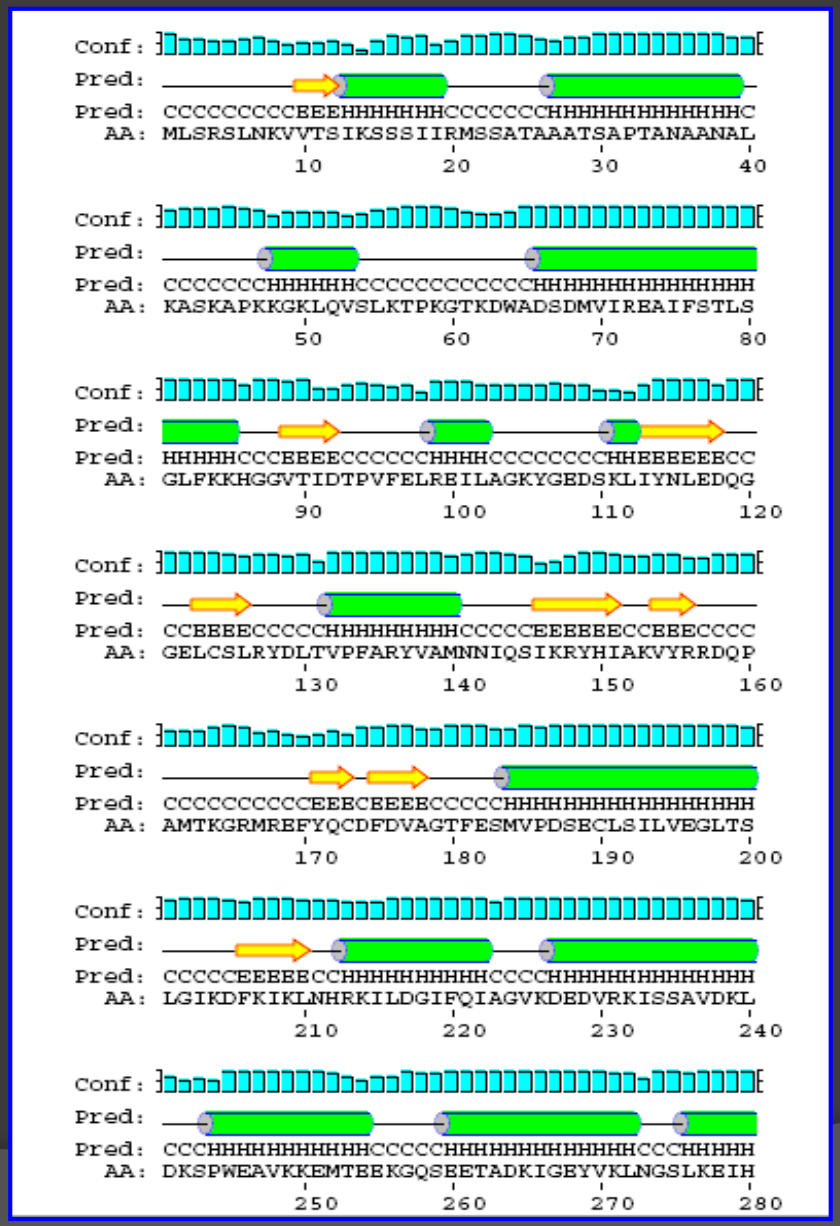
Conf: 997329
Pred: HHHHCC
AA: RLIKGL

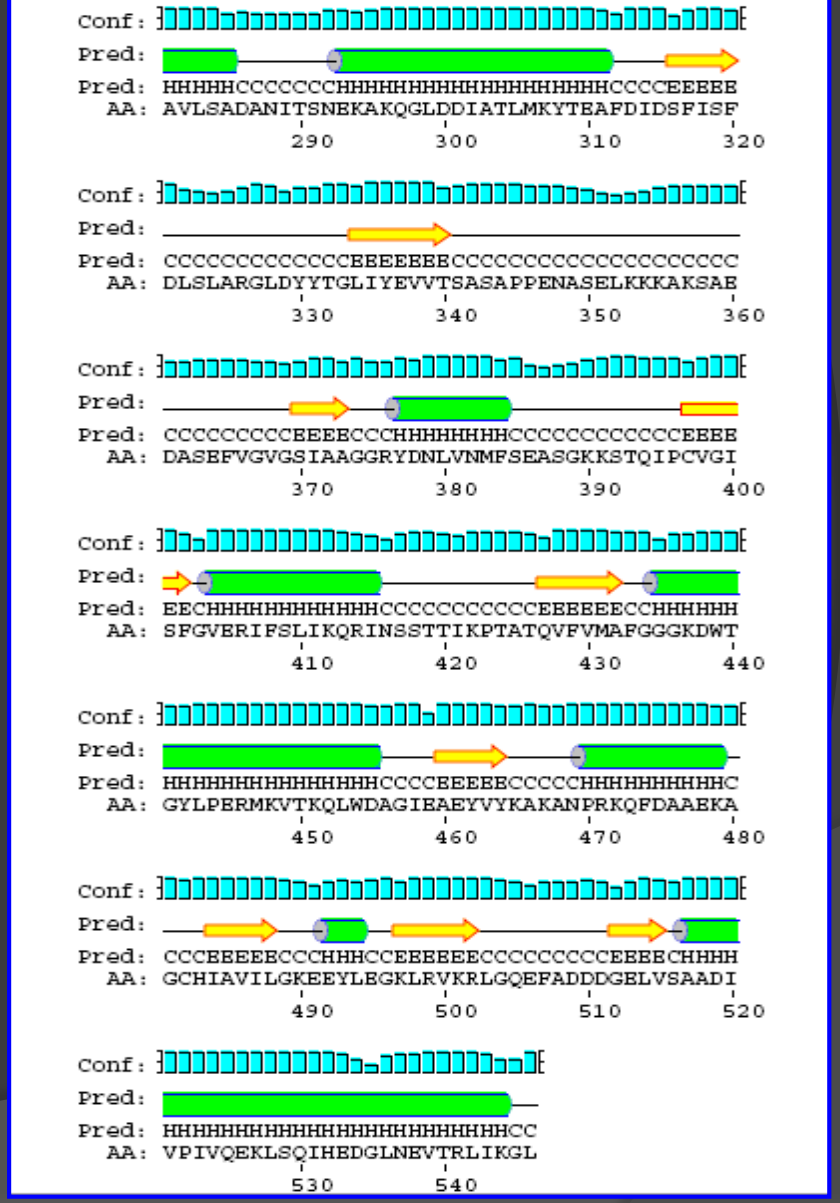
Calculate PostScript, PDF and JPEG graphical output for this result
using:
http://bioinf3.cs.ucl.ac.uk/cgi-bin/psipred/gra/nph-view2.cgi?id=082262e93452d467.psi


PSIPREDView Graphical Results - Microsoft Internet Explorer
File Modifica Visualizza Preferiti Strumenti ?
Indietro Avanti Termina Aggiorna Pagina iniziale Cerca
Indirizzo http://bioinf3.cs.ucl.ac.uk/cgi-bin/psipred/gra/nph-view2.cgi?id=082262e93452d
PSIP
Graphical output f
Liam J.
Bioinformat
University Colle
Click on
Format
PostScript File
PDF File
JPEG Page 1
JPEG Page 2
JPEG Page 3
Jones DT (1999) Protein secondary structure pr
McGuffin LJ, Bryson K, Jones DT (2000) Th
Adobe Reader - [082262e93452d467.psi.pdf]
File Edit View Document Tools Window Help
Save a Copy Search Select 100% Help
Adobe Reader 7.0
Conf:
Pred:
Pred: CCCCCCCCCBEEBHHHHHHHCCCCCCHHHHHHHHHHHHHC
AA: MLSRSLNKVVTSIKSSSIIRMSSATAAATSAPTANAANAL
10 20 30 40
Pred: CCCCCCCHHHHHHCCCCCCCCCCCCHHHHHHHHHHHHHHH
AA: KASKAPKKGKLQVSLKTPKGTKDWADSDMVIREAIFSTLS
50 60 70 80
Pred: HHHHHCCCEBEEBCCCCCCHHHHCCCCCCCCCHHEBEEEECC
AA: GLFKKHGGVTIDTPVFELREILAGKYGEDSKLIYNLEDQG
90 100 110 120
082262e93452d467 assunta 1 - Raccolta foto Windows
Aggiungi cartella alla raccolta File Correggi Informazioni Stampa Posta elettronica Masterizza Crea filmato Apri

Risultati predizione (versione grafica)







Jpred 3

Incorporating Jnet

What is Jpred?

JPred is not a single computer program, but is a web system. In the original JPred, a range of different secondary structure prediction algorithms were run and the results combined.

JNet

How does Jpred work?

Development of the JNet algorithm showed that this was more accurate than JPred, so this is now the default and only algorithm that is run.

The server runs in two modes; single sequence and multiple sequence.

1. **Multiple sequence** If you already have a set of aligned sequences you may submit them as either MSF format or BLC format, and the predictions will run. Your alignment will be modified so that it does not contain gaps in the first sequence. The first sequence should therefore always be your target sequence.
2. **Single sequence** For single sequences a multiple alignment is constructed. It is created by the PSI-BLAST algorithm with 3 iterations. Redundant sequences are removed and gaps that have appeared in the query sequence are removed along with the aligned positions in the sequences. The prediction algorithms are then run.

4formati

Raw/FASTA
MSF

BLC
BATCH

6metodi di predizione

Six different prediction methods (DSC [King & Sternberg, 1996], PHD [Rost & Sander, 1993], NNSSP [Salamov & Solovyev, 1995], PREDATOR [Frishman & Argos, ZPRED [Zvelebil *et al.*, 1987] and MULPRED (Barton, 1988, unpublished) are then run, and the results from each method are combined into a simple file format.

OUTPUT

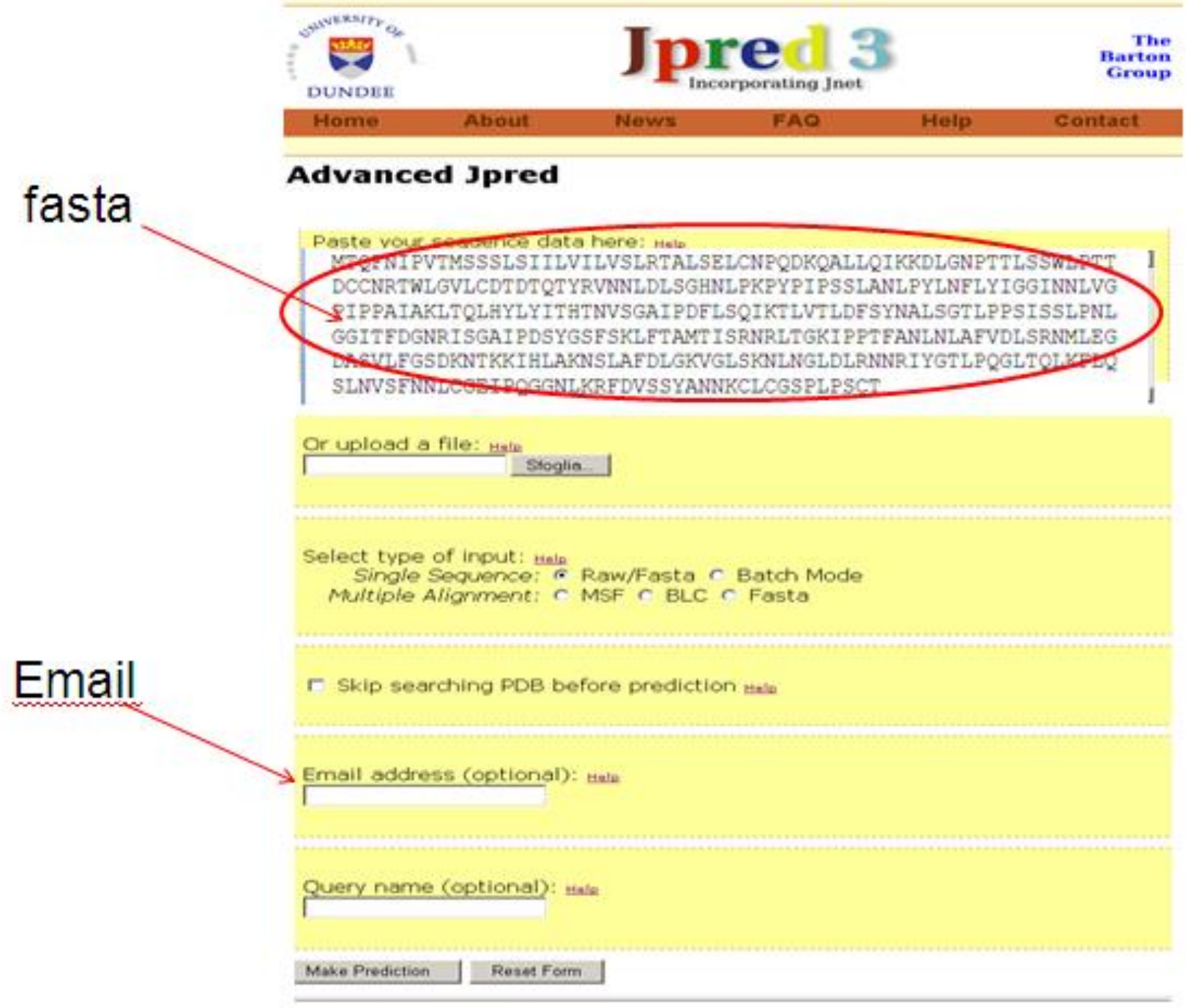
- Physico-chemical properties
- Solvent accessibility
- Prediction reliability
- Conservation number values

Publications: 2000

Cuff, J. A. and Barton, G. J. (2000), "Application of Multiple Sequence Alignment Profiles to Improve Protein Secondary Structure Prediction", *PROTEINS: Structure Function and Genetics*, **40**:502-511. [PubMed] (*EBI*)

This paper describes the JNet secondary structure prediction algorithm. JNet gives over 76% accuracy in a comprehensive blind test and also predicts burial of residues and assigns confidence levels to predictions. The JNet software is available for

2007 → 81,5%


fasta
Email
UNIVERSITY OF DUNDEE
Jpred 3
Incorporating Jnet
The Barton Group
Home
About
News
FAQ
Help
Contact
Advanced Jpred
Paste your sequence data here: Help
DCCNRTWLGVLCDTDTQTYRVNNLDLSGHNLPKPYPIPSSLANLPYLNFLYIGGINNLVG
GGITFDGNRISGAIPDSYGSFSKLFTAMTISRNRLTGKIPPTFANLNLAFVDLSRNMLEG
SLNVSFNNLCGEIPQGGNLKRFDVSSYANNKCLCGSPLPSCT
Or upload a file: Help
Sfoglia...
Select type of input: Help
Single Sequence: Raw/Fasta Batch Mode
Multiple Alignment: MSF BLC Fasta
Skip searching PDB before prediction Help
Email address (optional): Help
Query name (optional): Help
Make Prediction
Reset Form

Match found in PDB

You might want to reconsider the accuracy and what you might gain from secondary structure prediction, if close sequence homologues exist in the structural database.

If you still want to carry out a Jpred prediction click continue

Hits found

PDB	Chain	Description	Blast E-value
1acx	A	ACTINOXANTHIN	3e-27
1j48	B	Apoprotein of C1027	4e-25
1j48	A	Apoprotein of C1027	4e-25
1hzl	A	C-1027 APOPROTEIN	4e-25
1hzk	A	C-1027 APOPROTEIN	4e-25
2g0l	A	NEOCARZINOSTATIN	8e-07
2g0k	A	Neocarzinostatin	8e-07
1o5p	A	Neocarzinostatin	8e-07
1noa	A	NEOCARZINOSTATIN	8e-07
1nco	B	HOLO-NEOCARZINOSTATIN	8e-07
1nco	A	HOLO-NEOCARZINOSTATIN	8e-07
1j5i	A	PROTEIN (Apo-Neocarzinostatin)	8e-07
1j5h	A	Apo-Neocarzinostatin	8e-07
2mcm	A	MACROMOMYCIN	2e-06
2cbt	B	NEOCARZINOSTATIN	8e-06
2cbt	A	NEOCARZINOSTATIN	8e-06

PDB last upated on: 2008-05-21

Alignment of PDB hits to your sequence

```
>1acx_A mol:protein length:108  ACTINOXANTHIN
          Length = 108

 Score =  117 bits (293), Expect = 3e-27
 Identities = 58/58 (100%), Positives = 58/58 (100%)

Query: 1  APAFSVSPASGASDGQSVSVSVAAAGETYYIAQCAPVGGQDACNPATATSFTTDASGA 58
          APAFSVSPASGASDGQSVSVSVAAAGETYYIAQCAPVGGQDACNPATATSFTTDASGA
Sbjct: 1  APAFSVSPASGASDGQSVSVSVAAAGETYYIAQCAPVGGQDACNPATATSFTTDASGA 58


>1j48_B mol:protein length:110  Apoprotein of C1027
          Length = 110

 Score =  110 bits (275), Expect = 4e-25

Query: 1  APAFSVSPASGASDGQSVSVSV--AAAGETYYIAQCAPVGGQDACNPATATSFTTDASGA 58
          APAFSVSPASG SDGQSVSVSV  AAAGETYYIAQCAPVGGQDACNPATATSFTTDASGA
Sbjct: 1  APAFSVSPASGLSDGQSVSVSVSGAAAGETYYIAQCAPVGGQDACNPATATSFTTDASGA 60


>1hzl_A mol:protein length:110  C-1027 APOPROTEIN
          Length = 110

 Score =  110 bits (275), Expect = 4e-25
 Identities = 57/60 (95%), Positives = 57/60 (95%), Gaps = 2/60 (3%)

Query: 1  APAFSVSPASGASDGQSVSVSV--AAAGETYYIAQCAPVGGQDACNPATATSFTTDASGA 58
          APAFSVSPASG SDGQSVSVSV  AAAGETYYIAQCAPVGGQDACNPATATSFTTDASGA
Sbjct: 1  APAFSVSPASGLSDGQSVSVSVSGAAAGETYYIAQCAPVGGQDACNPATATSFTTDASGA 60


>1hzk_A mol:protein length:110  C-1027 APOPROTEIN
          Length = 110

 Score =  110 bits (275), Expect = 4e-25
 Identities = 57/60 (95%), Positives = 57/60 (95%), Gaps = 2/60 (3%)

Query: 1  APAFSVSPASGASDGQSVSVSV--AAAGETYYIAQCAPVGGQDACNPATATSFTTDASGA 58
          APAFSVSPASG SDGQSVSVSV  AAAGETYYIAQCAPVGGQDACNPATATSFTTDASGA
Sbjct: 1  APAFSVSPASGLSDGQSVSVSVSGAAAGETYYIAQCAPVGGQDACNPATATSFTTDASGA 60


>2g0l_A mol:protein length:122  NEOCARZINOSTATIN
          Length = 122

 Score = 49.7 bits (117), Expect = 8e-07
 Identities = 30/61 (49%), Positives = 38/61 (62%), Gaps = 3/61 (4%)

Query: 1  APAFSVSPASGASDGQSVSVSVAA--AGETYYIAQCAPVG-GQDACNPATATSFTTDASG 57
           AP  +V+P+SG SDG  V V+ A    AG  Y + QCA V  G  ACNPA  +S T DA+G
Sbjct: 11 APTATVTPSSGLSDGTVVKVAGAGLQAGTAYDVGQCAWVDTGVLACNPADFSSVTADANG 70

Query: 58 A 58
           +
Sbjct: 71 S 71


>2g0k_A mol:protein length:122  Neocarzinostatin
          Length = 122

 Score = 49.7 bits (117), Expect = 8e-07
 Identities = 30/61 (49%), Positives = 38/61 (62%), Gaps = 3/61 (4%)

Query: 1  APAFSVSPASGASDGQSVSVSVAA--AGETYYIAQCAPVG-GQDACNPATATSFTTDASG 57
           AP  +V+P+SG SDG  V V+ A    AG  Y + QCA V  G  ACNPA  +S T DA+G
Sbjct: 11 APTATVTPSSGLSDGTVVKVAGAGLQAGTAYDVGQCAWVDTGVLACNPADFSSVTADANG 70

Query: 58 A 58
           +
Sbjct: 71 S 71


>1nco_B mol:protein length:113  HOLO-NEOCARZINOSTATIN
          Length = 113

 Score = 49.7 bits (117), Expect = 8e-07
 Identities = 30/61 (49%), Positives = 38/61 (62%), Gaps = 3/61 (4%)

Query: 1  APAFSVSPASGASDGQSVSVSVAA--AGETYYIAQCAPVG-GQDACNPATATSFTTDASG 57
           AP  +V+P+SG SDG  V V+ A    AG  Y + QCA V  G  ACNPA  +S T DA+G
Sbjct: 2  APTATVTPSSGLSDGTVVKVAGAGLQAGTAYDVGQCAWVDTGVLACNPADFSSVTADANG 61

Query: 58 A 58
           +
Sbjct: 62 S 62


>1j5i_A mol:protein length:122  PROTEIN (Apo-Neocarzinostatin)
          Length = 122

 Score = 49.7 bits (117), Expect = 8e-07
 Identities = 30/61 (49%), Positives = 38/61 (62%), Gaps = 3/61 (4%)

Query: 1  APAFSVSPASGASDGQSVSVSVAA--AGETYYIAQCAPVG-GQDACNPATATSFTTDASG 57
           AP  +V+P+SG SDG  V V+ A    AG  Y + QCA V  G  ACNPA  +S T DA+G
Sbjct: 11 APTATVTPSSGLSDGTVVKVAGAGLQAGTAYDVGQCAWVDTGVLACNPADFSSVTADANG 70

Query: 58 A 58
           +
Sbjct: 71 S 71


>1j5h_A mol:protein length:122  Apo-Neocarzinostatin
          Length = 122

 Score = 49.7 bits (117), Expect = 8e-07
 Identities = 30/61 (49%), Positives = 38/61 (62%), Gaps = 3/61 (4%)

Query: 1  APAFSVSPASGASDGQSVSVSVAA--AGETYYIAQCAPVG-GQDACNPATATSFTTDASG 57
           AP  +V+P+SG SDG  V V+ A    AG  Y + QCA V  G  ACNPA  +S T DA+G
Sbjct: 11 APTATVTPSSGLSDGTVVKVAGAGLQAGTAYDVGQCAWVDTGVLACNPADFSSVTADANG 70

Query: 58 A 58
```

PDBsum entry: 1acx - Microsoft Internet Explorer

File Modifica Visualizza Preferiti Strumenti ?

Indietro Cerca Preferiti

Indirizzo http://www.ebi.ac.uk/pdbsum/1acx Vai Collegamenti

PDBsum

Top page | Protein | Clefts | Links

Antibacterial protein — PDB id: 1acx

Main view

Bottom view

Right view

Jmol

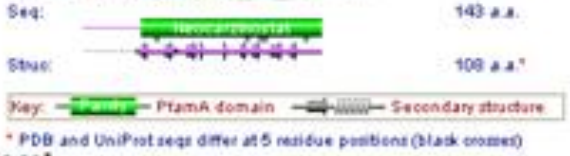
Contents

- Description
  - Header details
  - Header records
  - References
  - PROCHECK
- Protein chain
  - 108 a.a.

Tools

- Image Generation
- AstexViewer™@MSD-EBI
- Run PROCHECK
- Clefts Calculation

PDB id: 1acx
Name: Antibacterial protein
Title: Actinoxanthin structure at the atomic level (russian)
Structure: Actinoxanthin. Chain: a. Engineered: yes
Source: Streptomyces globisporus
UniProt: P01551 (ATXA_STRGL) [Pfam]



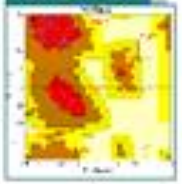
* PDB and UniProt seqs differ at 5 residue positions (black crosses)

Resolution: 2.00Å
R-factor: not given
Authors: V.Z.Pletnev,A.P.Kuzin
Key ref: . [PubMed id:] v.z.pletnev et al. (1982). Actinoxanthin Structure at the Atomic Level (Russian). *Bioorg.Khim.*, **8**, 1637.
Date: 17-Dec-82
Release date: 09-Mar-83

Quick links

- RCSB
- MSD
- SRS
- MMDB
- JenaLib
- OCA
- CATH
- SCOP
- FSSP
- HSSP
- PQS
- ProSAT
- Whatcheck

Procheck



Clefts

Surface

Jpred 3
Incorporating Inet


Jpred secondary structure prediction results
Full HTML View
http://www.compbio.dundee.ac.uk/~www-jpred/results/jp_mPVK
QUERY
UniRef90_A0ITA9
UniRef90_Q1I937
UniRef90_UP10000SF7DE1
UniRef90_A0UWJ4
UniRef90_C8TP72
UniRef90_A1DD43
OrigSeq
Jnet
jhmm
jpssm
Lupas 14
Lupas 21
Lupas 28
Jnet_25
Jnet_5
Jnet_0
Jnet Rel
Notes
Key:
Colour code for alignment:
Blue - Complete identity at a posi
Shades of red - The more red a position is,
conservation of chemical pr
Jnet - Final secondary structure p
jalign - Jnet alignment prediction
jhmm - Jnet hmm profile prediction
jpsm - Jnet PSIBLAST psm profile
Lupas - Lupas Coil prediction (wind
Note on coiled coil predictions - = less tha
c = between
C = greater
Jnet_25 - Jnet prediction of burial,
Jnet_5 - Jnet prediction of burial,
Jnet_0 - Jnet prediction of burial,
Jnet Rel - Jnet reliability of predict
Done
zotero
Jalview
http://www.compbio.dundee.ac.uk/~www-jpred/results/jp_mPVKBXv/jp_mPVKB
QUERY
UniRef90_A0ITA9
UniRef90_Q1I937
UniRef90_UP10000SF7DE1
UniRef90_A0UWJ4
UniRef90_C8TP72
UniRef90_A1DD43
Lupas_21
Lupas_14
Lupas_28
jnetpred
JNETCONF
JNETSOL25
JNETSOL5
JNETSOL0
JNETHMM
JNETPSSM
JNETJURY
Sequence 5 ID: UniRef90_A0UWJ4
PDF View
QUERY
UniRef90_A0ITA9
UniRef90_Q1I937
UniRef90_UP10000SF7DE1
UniRef90_A0UWJ4
UniRef90_C8TP72
UniRef90_A1DD43
lupas_21
lupas_14
lupas_28
jnet
conf
sol25
sol5
sol0
jhmm
jpsm

conclusion

They represent extended (E), helical (H) and other (-) types of secondary structure respectively. In the solvent accessibility predictions they represent buried (B) and exposed (-) for each of the 0%, 5% and 25% solvent accessibility cut-offs.

What do the colours mean in the Postscript/PDF output from Jpred?

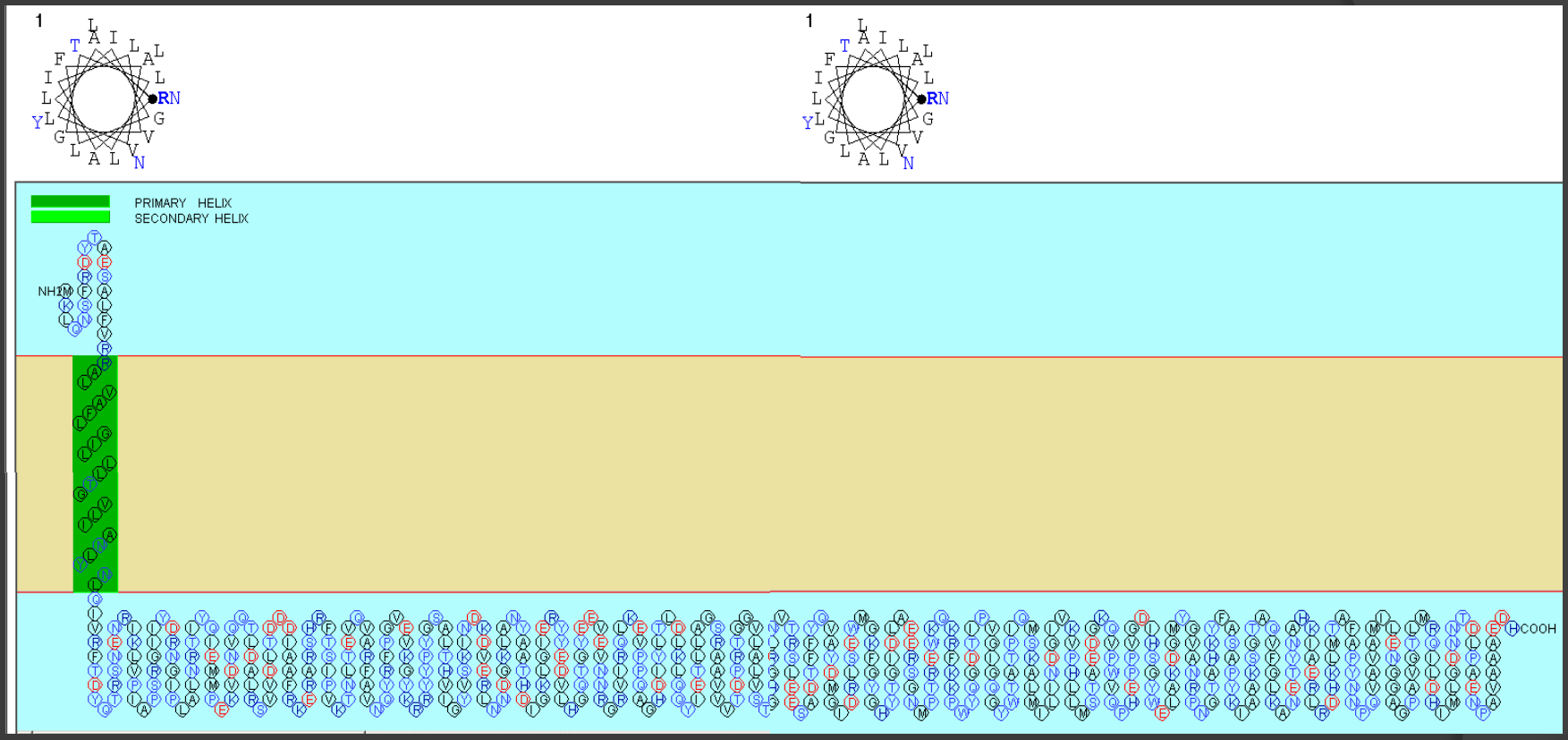
Character	Property
p	conserved polar
h	conserved hydrophobic
+	conserved positive charge
s	conserved small residues

Colour	Meaning
Pale blue	hydrophobic
Pale green	conserved polar
Small letters	small residues
Red	fully conserved
Blue text	Proline
Red text	Histidine
Boxed	Aliphatic (L, I or V)
Yellow	Cystine

Classification and Secondary Structure Prediction of Membrane Proteins

Mitaku Group
Department of Biotechnology
Tokyo University of Agriculture and Technology
mailto: sosui@proteome.bio.tuat.ac.jp
[Naogoya/ TUAT]

[Contents]



This amino acid sequence is of a MEMBRANE PROTEIN which have 1 transmembrane helix.

No.	N terminal	transmembrane regio	C terminal	type	length
1	20	RALVAFLGILLLTGVLIANLYNL	42	PRIMARY	23

TMpred output for unknown

[ISREC-Server] Date: Tue Nov 28 12:46:27 Europe/Zurich 2006

Sequence: MNG...APA, length: 348
Prediction parameters: TM-helix length between 17 and 33

1.) Possible transmembrane helices

The sequence positions in brackets denominate the core region.
Only scores above 500 are considered significant.

```
Inside to outside helices :   6 found
      from         to     score center
  37 (  41)  59 (  59)    2429     50
  74 (  74)  99 (  96)    1998     86
 117 ( 117) 139 ( 134)    1008    126
 153 ( 153) 175 ( 173)    2552    164
 205 ( 205) 222 ( 222)    2341    214
 253 ( 253) 274 ( 274)    3298    264

Outside to inside helices :   7 found
      from         to     score center
  37 (  39)  63 (  56)    2497     47
  75 (  78)  99 (  95)    1617     86
 115 ( 118) 140 ( 140)    1441    127
 153 ( 157) 176 ( 173)    1808    165
 203 ( 203) 221 ( 221)    2960    213
 253 ( 253) 273 ( 271)    2684    263
 286 ( 290) 309 ( 309)    1023    298
```

TMpred output for unknown

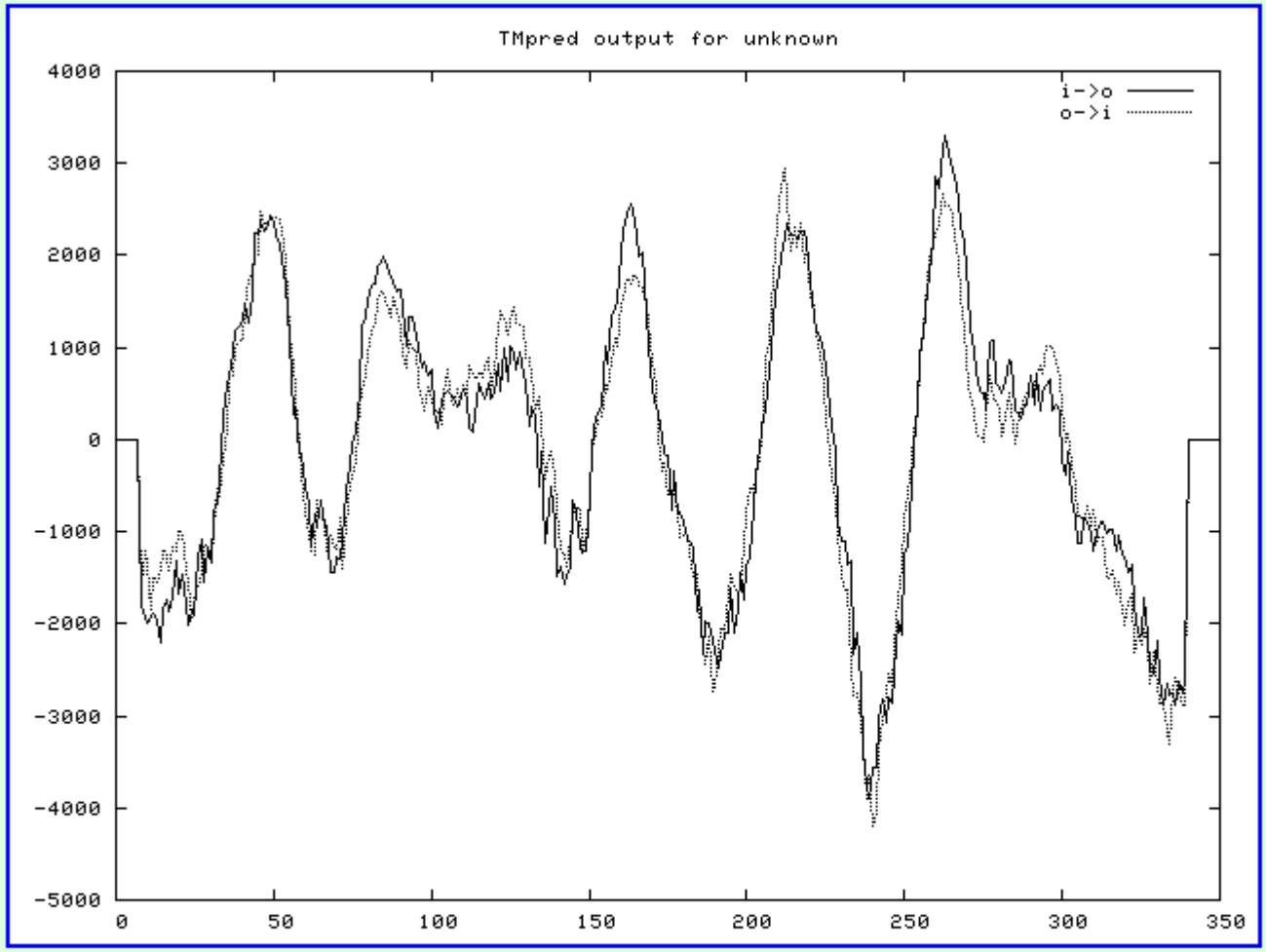
[ISREC-Server] Date: Tue Nov 28 12:46:27 Europe/Zurich 2006

Sequence: MNG...APA, length: 348
Prediction parameters: TM-helix length between 17 and 33

1.) Possible transmembrane helices

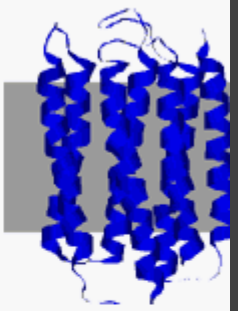
The sequence positions in brackets denominate the core region.
Only scores above 500 are considered significant.

```
Inside to outside helices :   6 found
      from         to     score center
  37 (  41)   59 (  59)    2429     50
  74 (  74)   99 (  96)    1998     86
 117 ( 117) 139 ( 134)    1008    126
 153 ( 153) 175 ( 173)    2552    164
 205 ( 205) 222 ( 222)    2341    214
 253 ( 253) 274 ( 274)    3298    264

Outside to inside helices :   7 found
      from         to     score center
  37 (  39)   63 (  56)    2497     47
  75 (  78)   99 (  95)    1617     86
 115 ( 118) 140 ( 140)    1441    127
 153 ( 157) 176 ( 173)    1808    165
 203 ( 203) 221 ( 221)    2960    213
 253 ( 253) 273 ( 271)    2684    263
 286 ( 290) 309 ( 309)    1023    298
```



You can get the prediction graphics shown above in one of the following formats:

- GIF-format
- Postscript-format
- numerical format



TMHMM Server v. 2.0

Prediction of transmembrane helices in proteins

Update Nov. 29 2001: Minor change to the html output.

NOTE: You can submit many proteins at once in one fasta file. Please limit each submission to at most 4000 proteins. Please tick the 'One line per protein' option. Please leave time between each large submission.

Instructions

SUBMISSION

Submission of a local file in FASTA format (HTML 3.0 or higher)

Browse...

OR by pasting sequence(s) in FASTA format:

```
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLYVTVQHKKLRT
PLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLGGEIALWSLVVLAIERYVVVC
KPMSNFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIPQGMQCSCGALYFTLKPEINNESFVIYMFVV
HFSIPLIVIFFCYGQLVFTVKEAAAQQQESATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQG
SDFGPIFMTIPAFFAKSSSVYNPVIYIMMNKQFRNCMLTTLCCGKNPLGDDEASTTVSKTETSQVAPA
```

Output format:

- (•) Extensive, with graphics
- ( ) Extensive, no graphics
- ( ) One line per protein

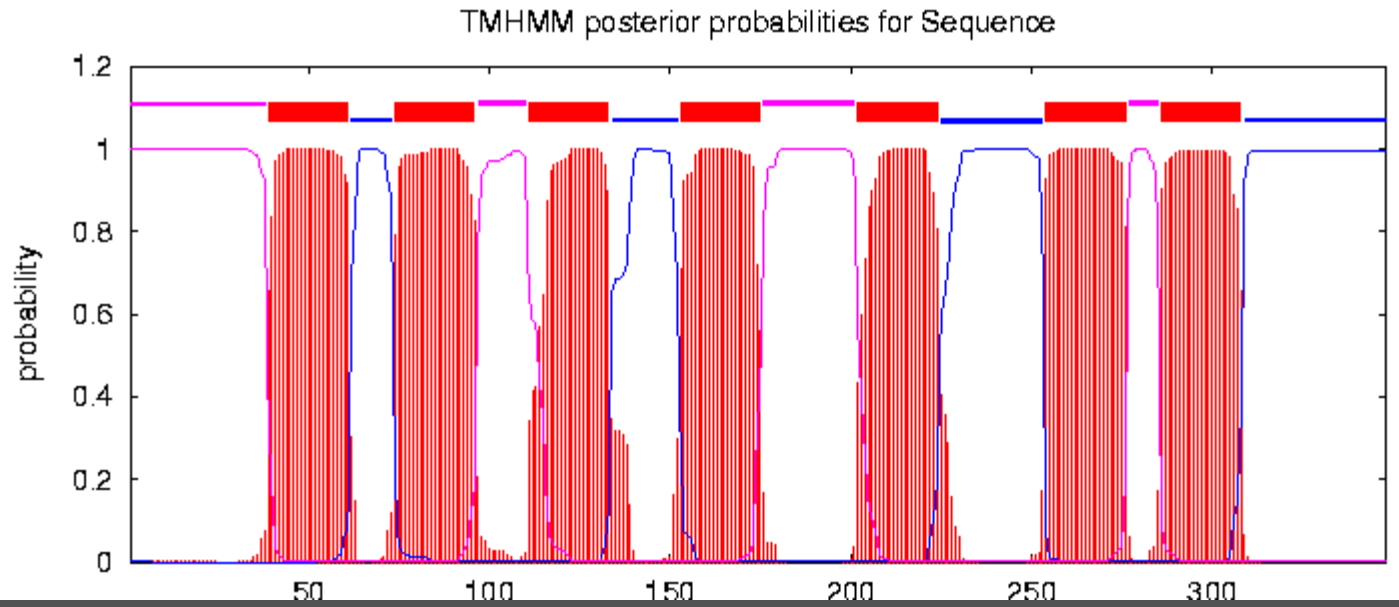
Other options:

- [ ] Use old model (version 1)

Submit Clear

```
# Sequence Length: 348
# Sequence Number of predicted TMHs:  7
# Sequence Exp number of AAs in TMHs: 157.8625
# Sequence Exp number, first 60 AAs:  21.49827
# Sequence Total prob of N-in:        0.00024
# Sequence POSSIBLE N-term signal sequence
```

Sequence	TMHMM2.0	outside	1	38
Sequence	TMHMM2.0	TMhelix	39	61
Sequence	TMHMM2.0	inside	62	73
Sequence	TMHMM2.0	TMhelix	74	96
Sequence	TMHMM2.0	outside	97	110
Sequence	TMHMM2.0	TMhelix	111	133
Sequence	TMHMM2.0	inside	134	152
Sequence	TMHMM2.0	TMhelix	153	175
Sequence	TMHMM2.0	outside	176	201
Sequence	TMHMM2.0	TMhelix	202	224
Sequence	TMHMM2.0	inside	225	253
Sequence	TMHMM2.0	TMhelix	254	276
Sequence	TMHMM2.0	outside	277	285
Sequence	TMHMM2.0	TMhelix	286	308
Sequence	TMHMM2.0	inside	309	348



Protein Classification

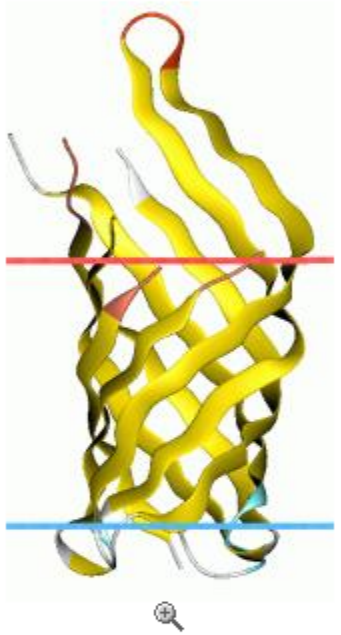
Types (3 types)
Classes (10 classes)
Superfamilies (168 superfamilies)
Families (238 families)
Species (237 species)
Localization (22 types)
All proteins in OPM (629 entries)

Protein Links

PDB Sum, PDB, MSD, SCOP, Swiss-Prot, Pfam, OCA, MMDB, FSSP, HSSP

1qjp » Outer membrane protein A (OMPA)

- *Type:* 1. Transmembrane (2 classes)
- *Class:* 1.2. Beta-barrel transmembrane (13 superfamilies)
- *Superfamily:* 1.2.01. OMPA-like (n=8,S=10) (2 families) 1.B.6 (TCDB)
- *Family:* 1.2.01.01. OMPA-like proteins (4 proteins) 1.B.6 (TCDB)
- *Species:* *Escherichia coli* (60 proteins)
- *Localization:* Bacterial gram-negative outer membrane (42 proteins)



3D view in Chime, Jmol or Webmol

Download Coordinates

Topology in Bacterial gram-negative outer membrane

1qjp » Outer membrane protein A (OMPA)	
Hydrophobic Thickness	24.9 ± 2.2 Å
Tilt Angle	8 ± 6°
$\Delta G_{transfer}$	-36.7 kcal/mol
Links to 1qjp	PDB Sum, SCOP, MSD, OCA, MMDB, Dali, HSSP
Topology	subunit A (N-terminus periplasmic)
Resolution	1.65 Å
Related PDB Sum	1bxw, 1q90, 2qe4

Perché interessa il profilo idrofobico/idrofilico: per predire anse tra elementi di Struttura secondaria i residui esposti e quelli sepolti regioni che attraversano La membrana i siti antigenici

Profilo idropatia ProtScale

SEQUENCE LENGTH: 633

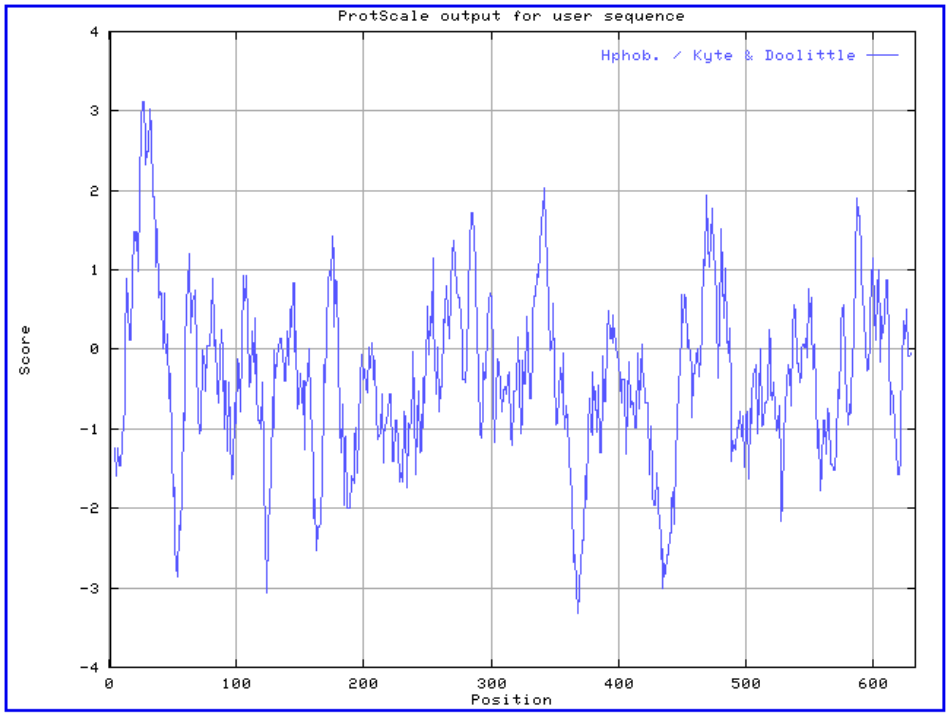
Using the scale **Hphob. / Kyte & Doolittle**, the individual values for the 20 amino acids are:

Ala: 1.800	Arg: -4.500	Asn: -3.500	Asp: -3.500	Cys: 2.500	Gln: -3.500
Glu: -3.500	Gly: -0.400	His: -3.200	Ile: 4.500	Leu: 3.800	Lys: -3.900
Met: 1.900	Phe: 2.800	Pro: -1.600	Ser: -0.800	Thr: -0.700	Trp: -0.900
Tyr: -1.300	Val: 4.200	Asx: -3.500	Glx: -3.500	Xaa: -0.490	

Weights for window positions 1,..,9, using **linear weight variation model**:

1	2	3	4	5	6	7	8	9
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
edge				center				edge

MIN: -3.322
MAX: 3.122



ScanProsite

The ScanProsite tool [Help / Commercial users] allows to scan protein sequence(s) (either from UniProt Knowledgebase (Swiss-Prot/TrEMBL) or PDB or provided by the user) for the occurrence of patterns, profiles and rules (motifs) stored in the PROSITE database, or to search protein database(s) for hits by specific motif(s) [Reference / Download ps_scan, the standalone version]. The program PRATT can be used to generate your own patterns. You may either:

- Enter one or more PROSITE accession numbers and/or patterns [1 by line] to search the UniProt Knowledgebase (Swiss-Prot/TrEMBL) and/or PDB databases, **OR**
- Enter one or more sequences [raw, Swiss-Prot or fasta format] and/or UniProt Knowledgebase (Swiss-Prot/TrEMBL) accession numbers and/or PDB accession numbers [1 by line] to be scanned with all patterns, profiles, rules in PROSITE, **OR**
- Fill in both fields to find all occurrences of specified motifs in specified sequences.

Protein(s) to be scanned:	**PROSITE pattern(s)/profile(s) to scan for:**
Enter one or more Swiss-Prot/TrEMBL accession number(s) [AC] (e.g. **P00747**) and/or sequence identifier(s) [ID] (e.g. **ENTK_HUMAN**), and/or PDB identifier, and/or paste **your own protein sequence(s)** in the box below: (leave this box blank to scan PROSITE entrie(s) against selected protein databases)	Enter one or more PROSITE accession number(s) (e.g. **PS50240**), and/or identifier(s) (e.g. **CHEB**), and/or type **your pattern(s)** in PROSITE format in the box below: (leave this box blank to scan sequence(s) against the entire PROSITE database)

Se si vuole testare la propria sequenza contro tutti i motivi di PROSITE
Oppure un codice di Swiss-Prot o TREMBL/PDB contro tutti i motivi di PROSITE
LATO GIALLO

Se si vuole testare il proprio PROSITE pattern o Profilo sul database
LATO AZZURRO

Si possono provare più sequenze contemporaneamente 8 o 16, stessa cosa per i motivi

CLASSIFICAZIONE delle proteine

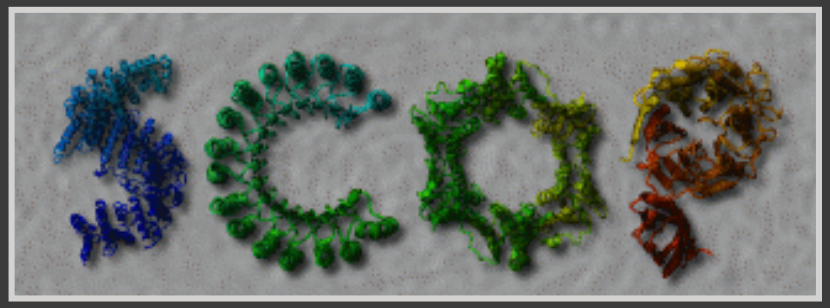
- Biochimica
  - Globulari
  - di Membrana
  - Fibrose
- Strutturale

Substrutture:

BANCHE DATI

- SCOP
- CATH

SCOP: Structural Classification Of Proteins (http://scop.mrc-lmb.cam.ac.uk/scop)



Classificazione e descrizione delle relazioni strutturali ed evolutive tra tutte le strutture proteiche conosciute

SCOP: Structural Classification Of Proteins (http://scop.mrc-lmb.cam.ac.uk/scop)

- Organizzato secondo dei livelli gerarchici:
  - Classe (strutturale)
  - Ripiegamento (strutturale)
  - Superfamiglia (evoluzionistico)
  - Famiglia (evoluzionistico)

SCOP

- Unità di categorizzazione: DOMINIO (*domain*)

SCOP: FAMIGLIA

★ Insiemi di domini omologhi → FAMIGLIA (*family*)

↓

Proteine con:

- Identità di sequenza ≥ 30%
- Funzioni e strutture molto simili

SCOP: FAMIGLIA

★ Insiemi di domini omologhi → FAMIGLIA (*family*)

↓

Proteine con:

Origine evolutiva comune

- Identità di sequenza ≥ 30%
- Funzioni e strutture molto simili

SCOP: SUPERFAMIGLIA

- Famiglie di proteine con:
  - Bassa identità di sequenza
  - Funzioni e strutture simili

SUPERFAMIGLIA
(*superfamily*)

SCOP: SUPERFAMIGLIA

- Famiglie di proteine con:
  - Bassa identità di sequenza
  - Funzioni e strutture simili

Origine evolutiva comune probabile

SUPERFAMIGLIA
(*superfamily*)

SCOP: RIPIEGAMENTO

- Due o più superfamiglie con una comune topologia di ripiegamento per un'ampia frazione della struttura

RIPIEGAMENTO
(*fold*)

SCOP: RIPIEGAMENTO

- Due o più superfamiglie con una comune topologia di ripiegamento per un'ampia frazione della struttura

Nessuna origine evolutiva comune

RIPIEGAMENTO
(*fold*)

SCOP: RIPIEGAMENTO

- Due o più superfamiglie con una comune topologia di ripiegamento per un'ampia frazione della struttura

Stesso ripiegamento: Ragioni chimico-fisiche

RIPIEGAMENTO (*fold*)

SCOP: CLASSE

- Ogni ripiegamento fa parte di una delle CLASSI (*Class*) generali.
  - Classi SCOP:
    - Tutto α
    - Tutto β
    - $\alpha+\beta$
    - α/β
    - *Multidomain protein*
    - *Membrane and Cell Surface Protein*
    - *Small protein*

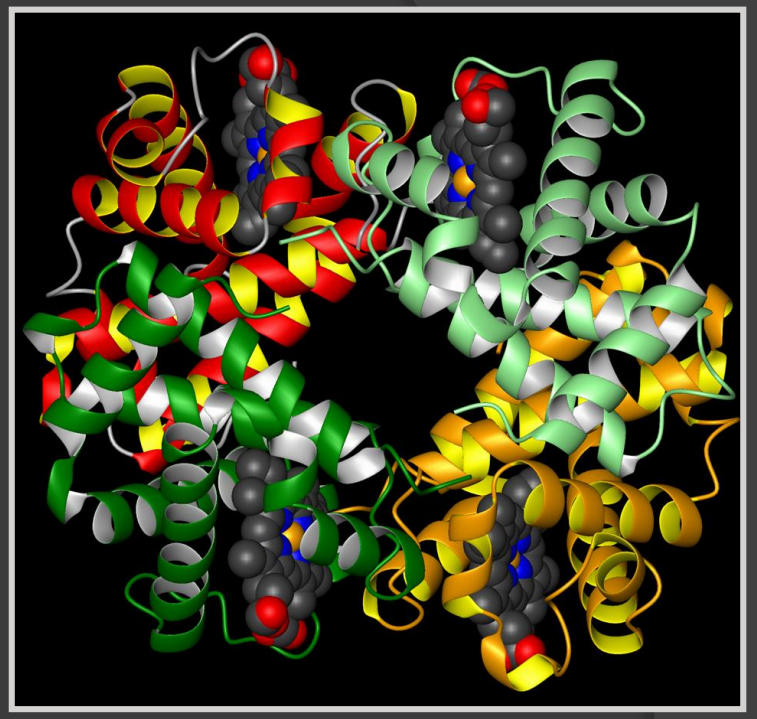
SCOP: Structural Classification Of Proteins

Scop Classification Statistics

SCOP: Structural Classification of Proteins. **1.71** release
27599 PDB Entries (18 Jan 2005). 75930 Domains. 1 Literature Reference
(excluding nucleic acids and theoretical models)

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	226	392	645
All beta proteins	149	300	594
Alpha and beta proteins (a/b)	134	221	661
Alpha and beta proteins (a+b)	286	424	753
Multi-domain proteins	48	48	64
Membrane and cell surface proteins	49	90	101
Small proteins	79	114	186
Total	971	1589	3004

SCOP



ESEMPIO:
Classificazione dell'emoglobina umana (catena a)

Protein: Hemoglobin, alpha-chain from Human (*Homo sapiens*)

Root: scop

- **Class**: All alpha proteins
- **Fold**: Globin-like *core: 6 helices; folded leaf, partly opened*
- **Superfamily**: Globin-like
- **Family**: Globins *Heme-binding protein*

Protein: Hemoglobin, alpha-chain
Species: Human (*Homo sapiens*)

CATH:

Class Architecture Topology and Homologous superfamily

(http://www.biochem.ucl.ac.uk/bsm/cath/)

- Database di Classificazione delle strutture proteiche presenti nella Protein Data Bank (PDB)

CATH:

Class Architecture Topology and Homologous superfamily

(http://www.biochem.ucl.ac.uk/bsm/cath/)

- Organizzato secondo dei livelli gerarchici:
  - Classe (class)
  - Architettura (architecture)
  - Topologia (topology)
  - Superfamiglia Omologa (homologous superfamily)

CATH: CLASSE

- Determinata dalla composizione della struttura secondaria e dell'impaccamento

CLASSE (*C-level*)

CATH: CLASSI

- *mainly* α
- *mainly* β
- $\alpha\beta$: α/β **e** $\alpha+\beta$
- Basso contenuto di struttura secondaria

CATH: ARCHITETTURA (*A-level*)

- Descrive la forma complessiva del dominio
- E' determinata dall'orientamento delle strutture secondarie nello spazio 3D, senza tener conto delle loro connessioni

Esempio: struttura a botte

CATH: TOPOLOGIA o famiglia di ripiegamento (*T-level*)

- Descrive il ripiegamento proteico tenendo conto dell'orientazione delle strutture secondarie e delle connessioni tra esse

CATH: SUPERFAMIGLIA OMOLOGA (*H-level*)

- Proteine con relazioni evoluzionistiche (antenato comune) = omologhe:
  - Similarità strutturale e/o funzionale
  - Alta identità di sequenza

Motif Scan

user: anonymous
log in

Protein Sequence Input
Enter a protein sequence in RAW or FASTA or Swiss-Prot format or a db:AC or db:ID identifier

Examples

```
MERTVLLLATVSLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQDILE
NGVKPLILRDCSVAGWLLGNPMCDEFINVPEWSYIVEKASPANDLCYPGN
LSRINHFEKIQIIPKSSWSNHDASSGVSSACPYLGRSSFFRNVVWLIKKN
NNTNQEDLLVLWGVHHPNDAAEQTKLYQNPTTYISVGTSTLNQRLVPEIA
RMEFFWTILKPNDAINFESNGNFIAPEYAYKIVKKGDSTIMKSELEYGNC
INSSMPFHNIHPLTIGECPKYVKSNRLVLATGLRNTPQRERRRKKRGLFG
QGMVDGWYGYHHSNEQGSCYSADKESTQKAIDGVTNKVNSIINKMNTQFE
RRIENLNKKMEDGFLDVWTYNAELLVLMENERTLDFHDSNVKNLYDKVRL
```

Clear input
Reset page

Motif scanning means finding all known motifs that occur in a sequence. This form lets you paste a protein sequence, select the collections of motifs to scan for, and launch the search. Some general **documentation** is available about the Prosite and Pfam collections of motifs. Another **document** deals with the interpretation of the match scores. You should consult the home pages of **Prosite** on ExPASy, **Pfam** and **InterPro** for additional information.

Warning: The scan might take a few minutes, thus if your proteins of interest are already in the sequence databases (see **list**), the **Query by Protein** form is much faster, and the **Protein Hub** provides a collection of tools that you might find useful.

Parameters

Database of motifs (**db description**)

- [x] PROSITE patterns
- [x] PROSITE patterns (frequent match producers)
- [x] PROSITE profiles
- [x] Prefile (more profiles)
- [ ] Na-channel profiles
- [x] HAMAP profiles
- [x] Pfam HMMs (local models)
- [x] Pfam HMMs (global models)

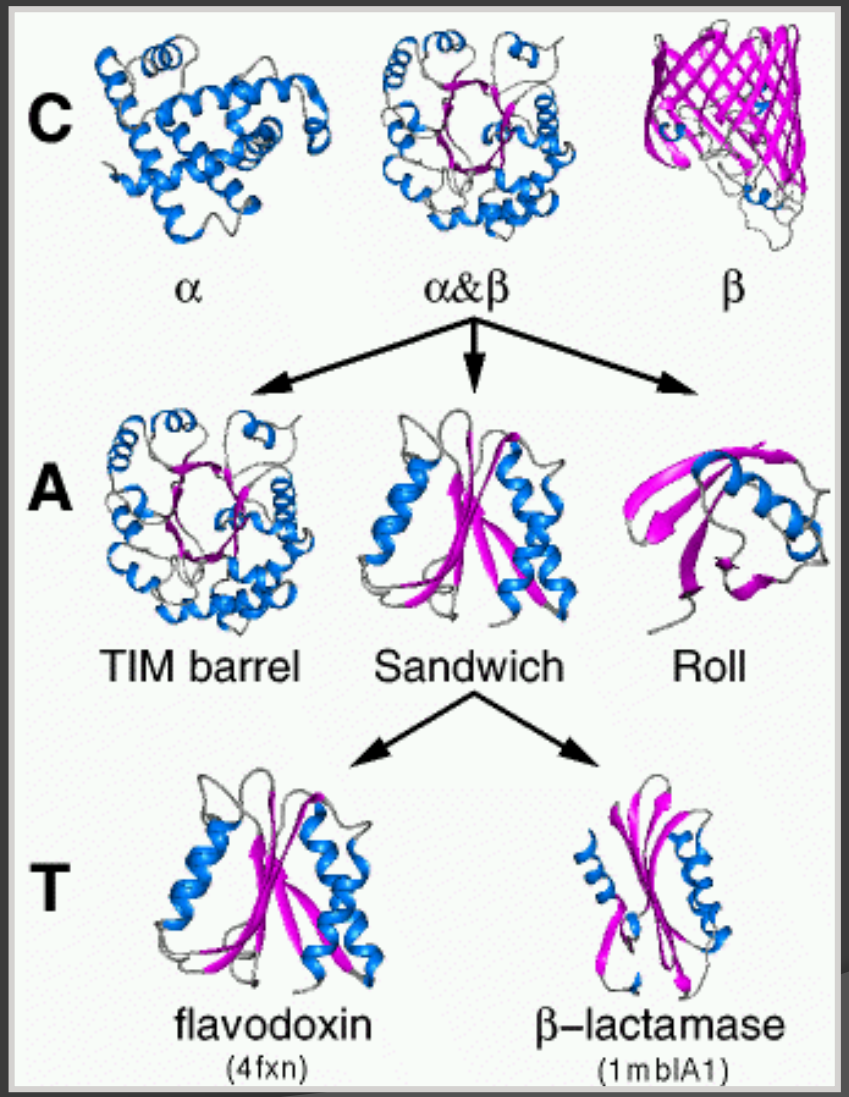
search

CATH: FAMIGLIA DI SEQUENZA (*S-level*)

- ★ Proteine di uno stesso *H-level*:
  - identità di sequenza ≥ 35%
  - similarità strutturale e/o funzionale

Level	Name	Sequence Identity Overlap
S	35%	80%
O	60%	80%
L	95%	80%
I	100%	80%

CATH: Esempio di livelli gerarchici



CATH:

Class Architecture Topology and Homologous superfamily

CATH v3.1.0

Version	3.1.0
Date	19-01-2007
Number of Domains	93885
Number of Chains	63453
Number of PDBs	30028

C	A	T	H	S	O	L	I	D
Mainly Alpha	5	305	652	1850	2329	3001	5587	19729
Mainly Beta	20	191	415	1860	2531	3846	6503	25537
Alpha Beta	14	496	922	3922	5303	6659	12998	47193
Few Secondary Structures	1	92	102	162	200	275	403	1426
Total	**40**	**1084**	**2091**	**7794**	**10363**	**13781**	**25491**	**93885**

CATH

ESEMPIO:

Classificazione dell'emoglobina umana (catena α)

CATH Domain 1spgA0

1spgA0

Classification

Class		1
Mainly Alpha		
Architecture		1.10
Orthogonal Bundle		
Topology		1.10.490
Globin-like		
Homologous Superfamily		1.10.490.10
Globins		
Sequence Family (S35)		1.10.490.10.3
Globins		
Non-identical (S95)		1.10.490.10.3.7
Globins		
Identical (S100)		1.10.490.10.3.7.1
Globins		

Motif Scan Results

user: anonymous
log in

Query Protein temporarily stored here.

Database of motifs PROSITE patterns, PROSITE patterns (frequent match producers), PROSITE profiles, Pfam HMMs (local models), Pfam HMMs (global models).

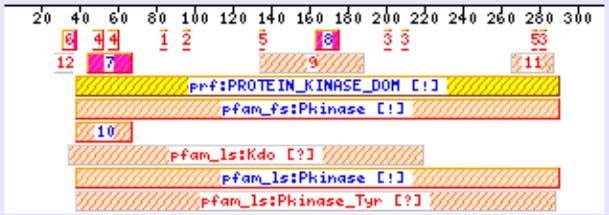
Reference: Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K & Bairoch A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.* **30**:235-238

searching PROSITE patterns
searching PROSITE patterns (frequent match producers)
searching PROSITE profiles
searching Pfam HMMs (local models)
searching Pfam HMMs (global models)
postprocessing

Summary

Original output pat, freq_pat, prf, pfam_fs, pfam_ls.

Matches map (features from query are above the ruler, matches of the motif scan are below the ruler)



Legends: 1, freq_pat:ASN_GLYCOSYLATION [?]; 2, freq_pat:CAMP_PHOSPHO_SITE [?]; 3, freq_pat:CK2_PHOSPHO_SITE [?]; 4, freq_pat:MYRISTYL [?]; 5, freq_pat:PKC_PHOSPHO_SITE [?]; 6, freq_pat:TYR_PHOSPHO_SITE [?]; 7, pat:PROTEIN_KINASE_ATP [!]; 8, pat:PROTEIN_KINASE_ST [!]; 9, pfam_fs:APH [?]; 10, pfam_fs:Pkinase_Tyr [!]; 11, pfam_fs:Pkinase_Tyr [?]; 12, pfam_ls:Involucrin [?].

FT	MYHIT	82	85	freq_pat:ASN_GLYCOSYLATION [?]
FT	MYHIT	94	97	freq_pat:CAMP_PHOSPHO_SITE [?]
FT	MYHIT	199	202	freq_pat:CK2_PHOSPHO_SITE [?]
FT	MYHIT	208	211	freq_pat:CK2_PHOSPHO_SITE [?]
FT	MYHIT	280	283	freq_pat:CK2_PHOSPHO_SITE [?]
FT	MYHIT	47	52	freq_pat:MYRISTYL [?]
FT	MYHIT	55	60	freq_pat:MYRISTYL [?]
FT	MYHIT	134	136	freq_pat:PKC_PHOSPHO_SITE [?]

Status: ?
pos.: **31-38**

freq_pat:TYR_PHOSPHO_SITE
Tyrosine kinase phosphorylation site.
[entry]

Legends: 1, phosphorylation.

^ image ^

Status: !
pos.: **44-67**

pat:PROTEIN_KINASE_ATP
Protein kinases ATP-binding region signature.
[entry]

^ image ^

Status: !
pos.: **163-175**

pat:PROTEIN_KINASE_ST
Serine/Threonine protein kinases active-site signature.
[entry]

Match Status Code	
!	A strong match: it is very unlikely that this match is a false positive.
R	Rescued match: despite the low score, it is considered to be a strong match. This concerns primarily domains known to be repeated and that are unlikely to occur as a single copy in a protein.
?	Questionnable or weak match: determining the true or false negative status of this match requires additional biological evidences.
!!	Strong match for a family-specific motif: it is very unlikely that this match is a false positive, in addition it is very likely that this match belongs to the targetted sub-family.
?!	Accepted match for a family-specific motif: it is very unlikely that this match is a false positive for the motif, but determining its family assignment requires additional biological evidences.
??	Questionnable or weak match for a family-specific motif: determining the true or false negative status of this match requires additional biological evidences.
NA	Not Available. The above interpretation rules make no sense (e.g., for a low-complexity region).

Prosite pattern e profile permette di caratterizzare la funzione di una proteina non caratterizzata database di siti e pattern biologicamente significativi.

Pfam è un database in cui vengono raccolte famiglie e motivi di proteine

PfamHMM

InterPro database di famiglie di proteine

HAMAP profile raccolta di famiglie di proteine ortologhe microbiche

TIGRfam raccolta di famiglie di proteine

NiceSite View of PROSITE: PS00128

General information about the entry	
Entry name	**LACTALBUMIN_LYSOZYME**
Accession number	**PS00128**
Entry type	PATTERN
Date	APR-1990 (CREATED); NOV-1997 (DATA UPDATE); OCT-2006 (INFO UPDATE).
PROSITE documentation	PDOC00119
Name and characterization of the entry	
Description	Alpha-lactalbumin / lysozyme C signature.
Pattern	C-x(3)-C-x(2)-[LMF]-x(3)-[DEN]-[LI]-x(5)-C.

Numerical results

- UniProtKB/Swiss-Prot release number: **51.1**, total number of sequence entries in that release: **241365**.
- Total number of hits in UniProtKB/Swiss-Prot: **129 hits in 129 different sequences**
- Number of hits on proteins that are known to belong to the set under consideration: **128 hits in 128 different sequences**
- Number of hits on proteins that could potentially belong to the set under consideration: **0 hits in 0 different sequences**
- Number of false hits (on unrelated proteins): **1 hits in 1 different sequences**
- Number of known missed hits: **2**
- Number of partial sequences which belong to the set under consideration, but which are not hit by the pattern or profile because they are partial (fragment) sequences: **7**
- Precision (true hits / (true hits + false positives)): **99.22 %**
- Recall (true hits / (true hits + false negatives)): **98.46 %**

Comments

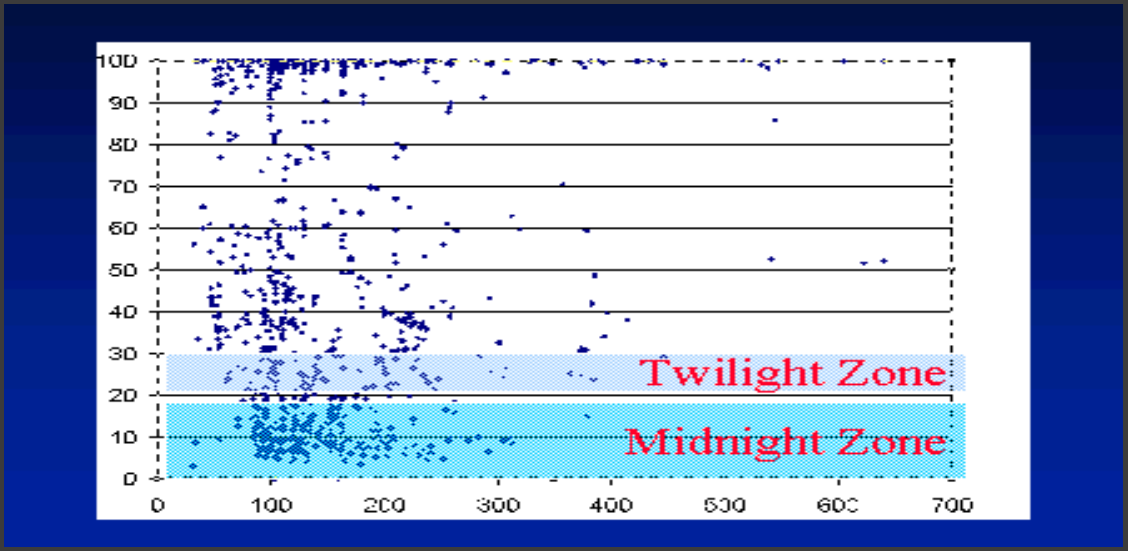
- Taxonomic range: **Eukaryotes**

Effetto IDROFOBICO
Evento fondamentale del folding

- Segregazione delle catene laterali idrofobiche:

1. Riduzione del numero delle conformazioni possibili per ingombro sterico;

2. I gruppi –NH e –CO dei residui segregati, trovandosi vicino, formano legami ad H → Formazione di elementi di struttura secondaria



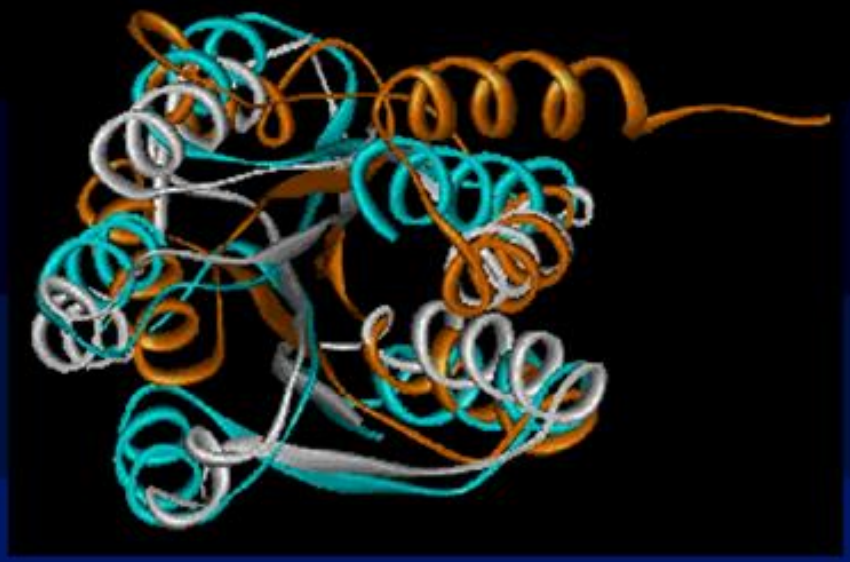
Zona crepuscolare

Sono state raccolte in modo casuale una certa quantità di sequenze presenti nel PDB e le loro sequenze sono state confrontate tra di loro.

Zona Certa Sequenze con similarità superiore al 40% per il 95% dei casi sono omologhe.

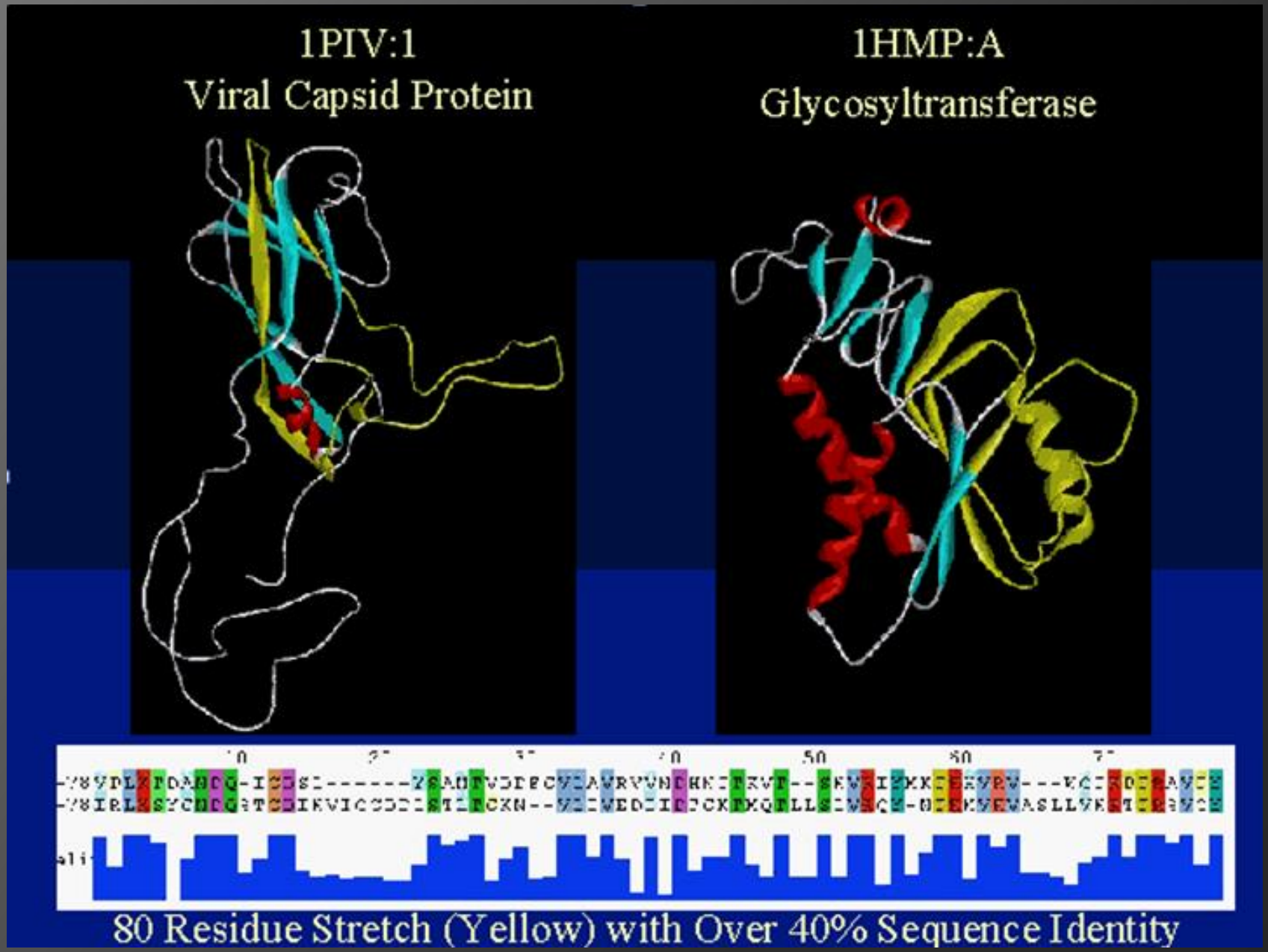
Zona ambigua In cui la similarità è tra il 20% - 30% solo il 10% dei casi le proteine sono omologhe

C'è poi una zona in cui la percentuale di similarità è inferiore al 20% ed in cui non si trovano pochissime sequenze omologhe



1ymv 1fla 1pdo

Alpha/beta proteins characterized as different superfamilies


1PIV:1
Viral Capsid Protein
1HMP:A
Glycosyltransferase
80 Residue Stretch (Yellow) with Over 40% Sequence Identity

Dall'analisi del PDB

Ci sono circa 1000 famiglie di proteine composte da membri di proteine che mostrano una certa similarità di sequenza

Una nuova sequenza mostrerà quasi sicuramente similarità di sequenza con altre e componenti strutturali simili.

C'è un limitato numero di ripiegamenti, nuove sequenze di solito mostrano ripiegamenti similari con strutture già note

Sequenze completamente diverse possono ripiegarsi in strutture simili

Ci sono 3 metodi di approccio per predire la struttura proteica

Metodi Ab-initio

Modeling comparativo

Fold Recognition

La predizione è basata su un calcolo computazionale proveniente dalla posizione di ogni atomo nello spazio e le sue relazioni chimico-fisiche con altri atomi

Teoreticamente possibile

Praticamente poco possibile

SWISS-MODEL

An Automated Comparative Protein Modelling Server

SIB - Biozentrum Basel site provided by:

BIOZENTRUM SIB

SWISS-MODEL is a fully automated protein structure homology-modeling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer). The purpose of this server is to make Protein Modelling accessible to all biochemists and molecular biologists World Wide.

The present version of the server is 3.5 and is under constant improvement and debugging. In order to help us refine the sequence analysis and modelling algorithms, please report of possible bugs and problems with the modelling procedure.

SWISS-MODEL was initiated in 1993 by Manuel Peitsch, and is now being further developed within the SIB - Swiss Institute of Bioinformatics in collaboration between Torsten Schwede at the Structural Bioinformatics Group, Biozentrum (University of Basel) and Nicolas Guex at GlaxoSmithKline.
The computational resources for the SWISS-MODEL server are provided in collaboration by the Biozentrum (University Basel) and the Advanced Biomedical Computing Center (NCI Frederick, USA).

Disclaimer

The result of any modelling procedure is NON-EXPERIMENTAL and MUST be considered with care. This is especially true since there is no human intervention during model building. Carefully read the header section of the files to know what templates and alignments were used during the model building process.

SWISS MODEL

E' un server automatizzato e comparativo per il modelling proteico nato nel 1993 da Manuel Peitsch

Reference
Arnold K., Bordoli L., Kopp J., and Schwede T. (2006). The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling. Bioinformatics, 22,195-201.
Kopp J. and Schwede T. (2004) The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models Nucleic Acids Research 32, D230-D234.
Schwede T, Kopp J, Guex N, and Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Research 31: 3381-3385.
Guex, N. and Peitsch, M. C. (1997) SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modelling. Electrophoresis 18: 2714-2723.
Peitsch, M. C. (1995) Protein modeling by E-mail Bio/Technology 13: 658-660.

ATTENZIONE!!

The result of any modelling procedure is NON-EXPERIMENTAL and MUST be considered with care.This is especially true since there is in human intervention during model building. Carefully read the header section of the files to know what templates and alignments were used during the model building process

Step 1

Serch for suitable template throught BLASTp2.

Database : ExNRL-3D

Step 2

Check sequence identity with target

Programma : SIM

Step 3

Generate models with ProModII

Programma: ProModII
Database: ExPDB

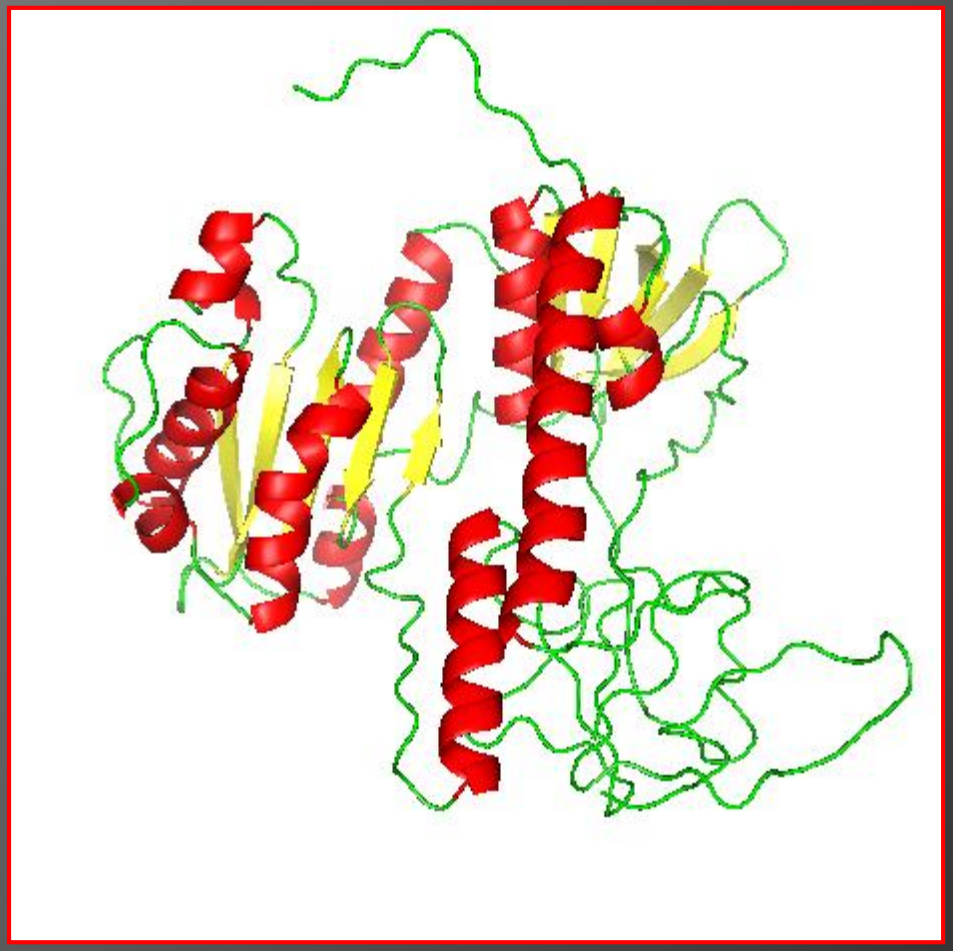
ProModII

At least one known 3D-structure of a related protein.
Good quality sequence alignement; the reliability of the model is determined by the degree of sequence identity

1. Superposition of related 3D-structure
2. Generation of a multiple a alignement with the sequence to be modelled.
3. Generation of a framework for the new sequence.
4. Rebuild lacking loops.
5. Complete and correct backbone.
6. Correct and rebuild side chains.
7. Verify model strucure quality and check packing.
8. Refine structure by energy minimisation and Molecular Dynamics.

Step 4

Energy minimisation with Gromos 96



Model info:

modelled residue range:

18 to 404

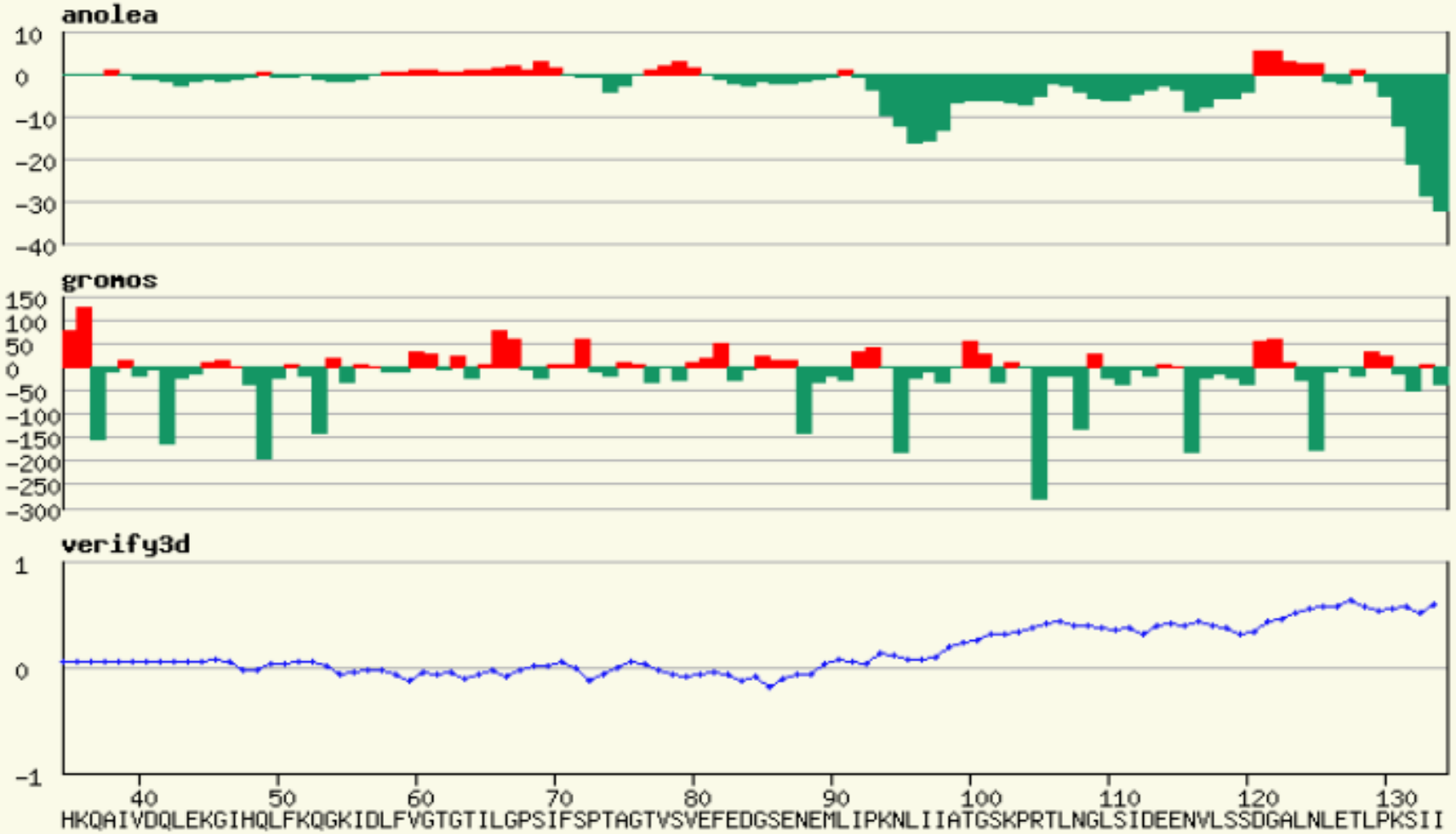
based on template:

2yquB (1.70 Å)

Sequence Identity [%]: 41

Evalue:

5.02e-65


anolea
10
0
-10
-20
-30
-40
gromos
150
100
50
0
-50
-100
-150
-200
-250
-300
verify3d
1
0
-1
40
50
60
70
80
90
100
110
120
130
HKQAIVDQLEKGIHQLFKQGKIDLFVGTGTILGPSIFSPTAGTVSVEFEDGSENEMLIPKNLIIATGSKPRTLNGLSIDEENVLSSDGALNLETLPKSII
anolea
10
0
-10
-20
-30
-40

Pôle BioInformatique Lyonnais

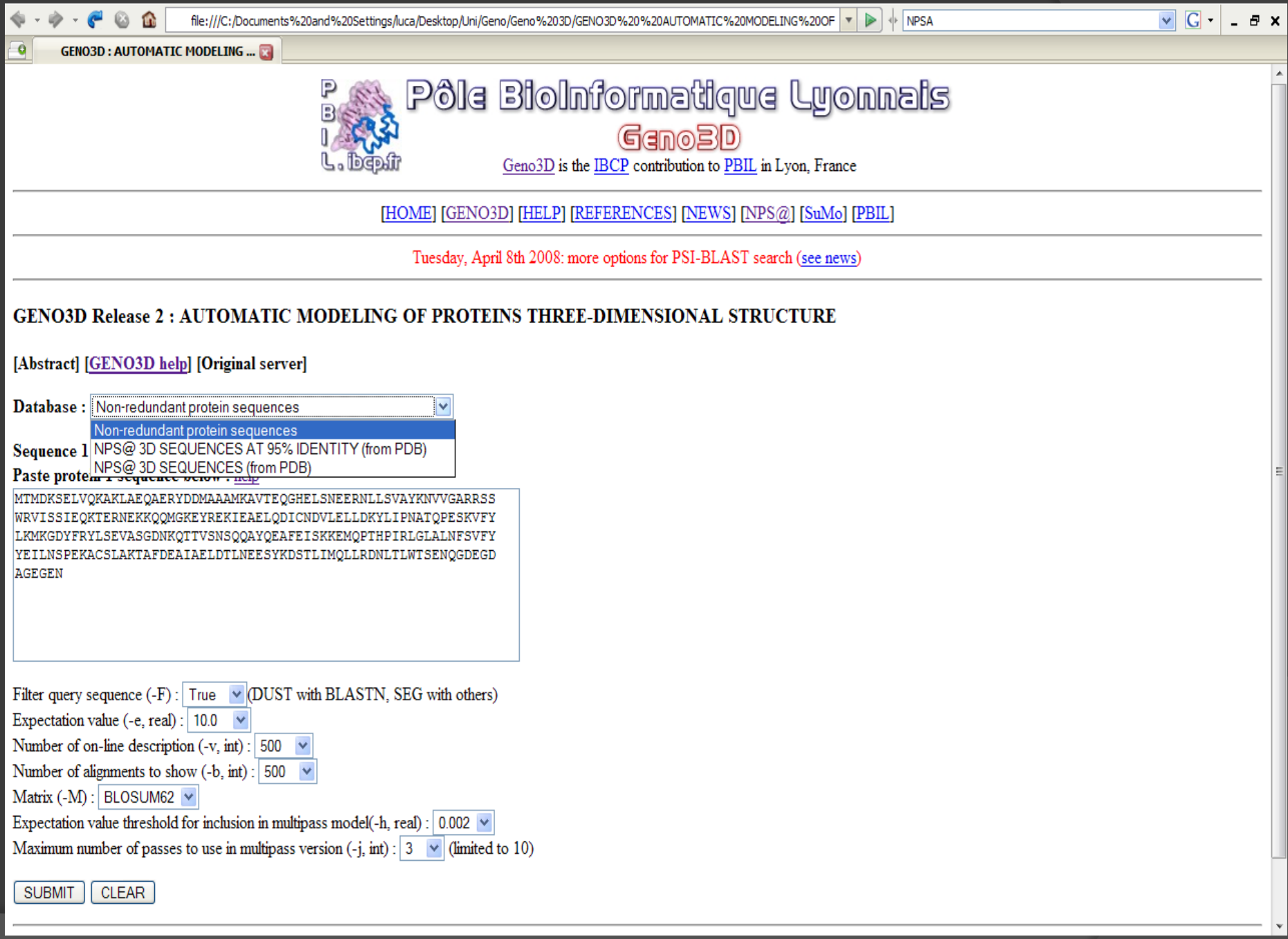
Geno3D

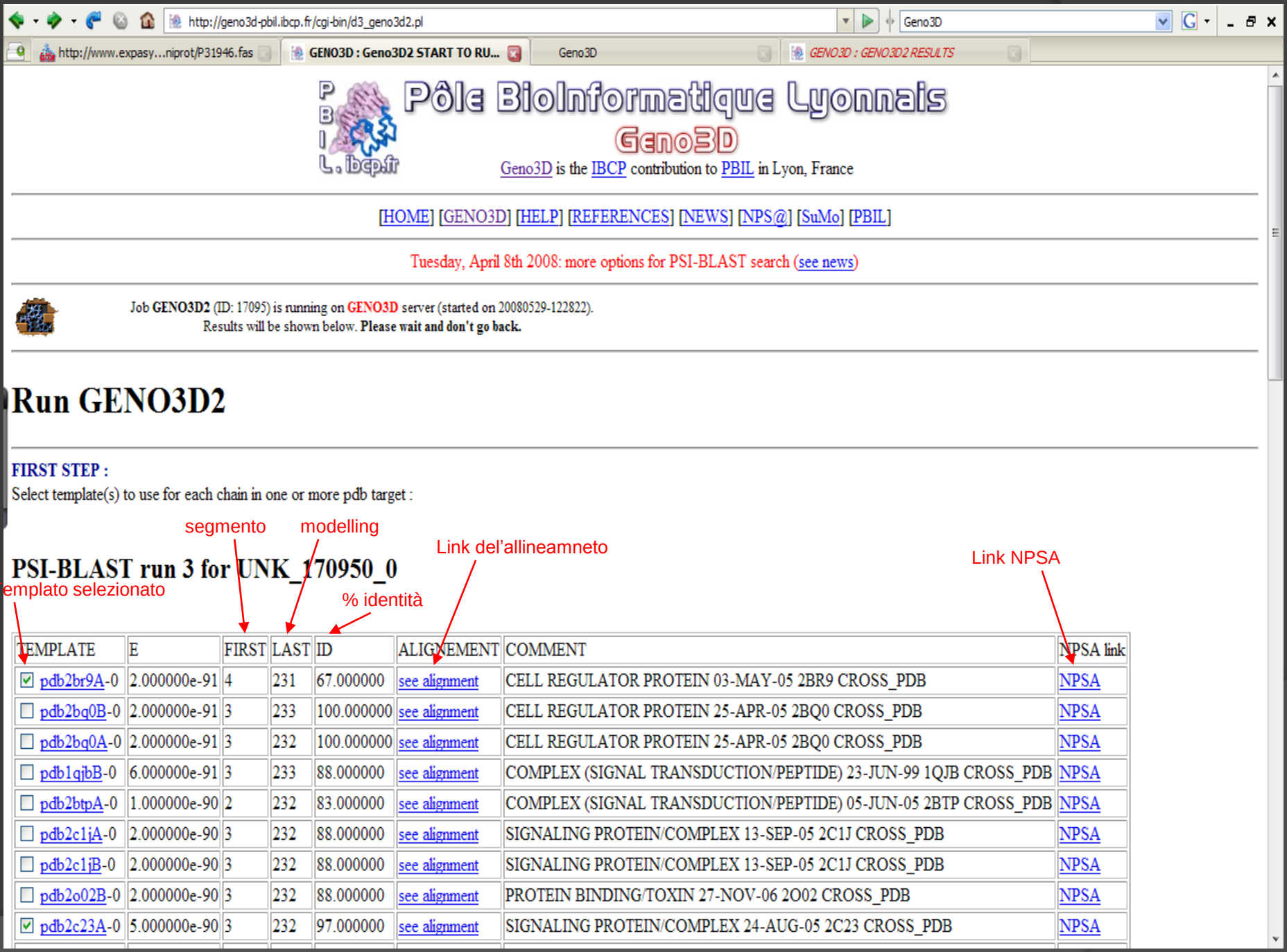
Geno3D is the IBCP contribution to PBIL in Lyon, France

- Server utilizzabile per generare modelli proteici 3D non sperimentali, tramite modelling per omologia;
- E' possibile utilizzare modelli con basso grado di identità (fino al 20%);
- Genera modelli fino a 500 amminoacidi;
- E’ possibile fare allineamenti multipli (fino a 3 contemporaneamente).

- L'operatore fornisce la query che viene confrontata e allineata, tramite PSI-BLAST, con tutte le voci presenti nella banca dati di PDB;
- L'utente seleziona le voci che vuole siano utilizzate nella modellizzazione molecolare.
- Il server calcola la predizione di struttura della query basandosi su ciascun templato e ne calcola la percentuale di accordo.
- In caso di selezione di più templati è visualizzato lo scarto quadratico medio tra il carbonio α e i vari templati.
- Le restrizioni degli angoli e le distanze tra gli atomi vengono calcolati a partire dall'allineamento con i templati 3D, per poi essere applicati sulla sequenza query. Per i gap vengono utilizzati calcoli statistici.
- Il software utilizzato per calcolare le restrizioni è CNS.

- Alla fine del processo di modellizzazione molecolare si riceve una e-mail in cui è fornito un indirizzo internet dove sono presenti i risultati (disponibili per 7 gg). E' possibile utilizzare questi risultati on line o scaricarli sul proprio pc in formato .archive.tar.gz
- In output vengono rilasciati i modelli 3D che soddisfano le restrizioni nel migliore dei modi.
- Vi possono essere porzioni che non vengono risolte basandosi sui templati; queste possono essere particolari domini che vengono ricostruiti usando templati di proteine omologhe.

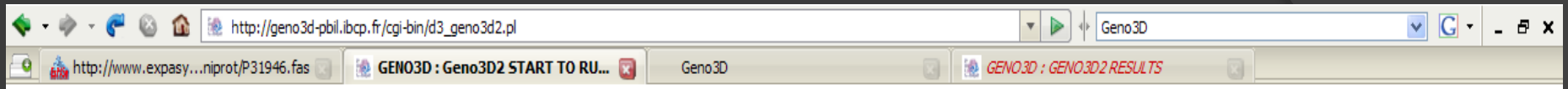

file:///C:/Documents%20and%20Settings/luca/Desktop/Uni/Geno/Geno%203D/GENO3D%20%20AUTOMATIC%20MODELING%20OF
NPSA
GENO3D : AUTOMATIC MODELING ...
Pôle Bioinformatique Lyonnais
Geno3D
Geno3D is the IBCP contribution to PBIL in Lyon, France
[HOME] [GENO3D] [HELP] [REFERENCES] [NEWS] [NPS@] [SuMo] [PBIL]
Tuesday, April 8th 2008: more options for PSI-BLAST search (see news)
GENO3D Release 2 : AUTOMATIC MODELING OF PROTEINS THREE-DIMENSIONAL STRUCTURE
[Abstract] [GENO3D help] [Original server]
Database :
Non-redundant protein sequences
Non-redundant protein sequences
NPS@ 3D SEQUENCES AT 95% IDENTITY (from PDB)
NPS@ 3D SEQUENCES (from PDB)
Sequence 1
Paste protein sequence below : help
MTMDKSELVQKAKLAEQAERYDDMAAAMKAVTEQGHELSNEERNLLSVAYKNVVGARRSS
WRVISSIEQKTERNEKKQQMGKEYREKIEAELQDICNDVLELLDKYLIPNATQPESKVFY
LKMKGDYFRYLSEVASGDNKQTTVSNSQQAYQEAFEISKKEMQPTHPIRLGLALNFSVFY
YEILNSPEKACSLAKTAFDEAIAELDTLNEESYKDSTLIMQLLRDNLTLWTSENQGDEGD
AGEGEN
Filter query sequence (-F) : True (DUST with BLASTN, SEG with others)
Expectation value (-e, real) : 10.0
Number of on-line description (-v, int) : 500
Number of alignments to show (-b, int) : 500
Matrix (-M) : BLOSUM62
Expectation value threshold for inclusion in multipass model(-h, real) : 0.002
Maximum number of passes to use in multipass version (-j, int) : 3 (limited to 10)
SUBMIT
CLEAR


http://geno3d-pbil.ibcp.fr/cgi-bin/d3_geno3d2.pl
Geno3D
http://www.expasy...niprot/P31946.fas
GENO3D : Geno3D2 START TO RU...
Geno3D
GENO3D : GENO3D2 RESULTS
Pôle Bioinformatique Lyonnais
Geno3D
Geno3D is the IBCP contribution to PBIL in Lyon, France
[HOME] [GENO3D] [HELP] [REFERENCES] [NEWS] [NPS@] [SuMo] [PBIL]
Tuesday, April 8th 2008: more options for PSI-BLAST search (see news)
Job GENO3D2 (ID: 17095) is running on GENO3D server (started on 20080529-122822).
Results will be shown below. Please wait and don't go back.
Run GENO3D2
FIRST STEP :
Select template(s) to use for each chain in one or more pdb target :
segmento
modelling
Link del'allineamneto
Link NPSA
PSI-BLAST run 3 for UNK_170950_0
Templato selezionato
% identità
TEMPLATE E FIRST LAST ID ALIGNEMENT COMMENT NPSA link
pdb2br9A-0 2.000000e-91 4 231 67.000000 see alignment CELL REGULATOR PROTEIN 03-MAY-05 2BR9 CROSS_PDB NPSA
pdb2bq0B-0 2.000000e-91 3 233 100.000000 see alignment CELL REGULATOR PROTEIN 25-APR-05 2BQ0 CROSS_PDB NPSA
pdb2bq0A-0 2.000000e-91 3 232 100.000000 see alignment CELL REGULATOR PROTEIN 25-APR-05 2BQ0 CROSS_PDB NPSA
pdb1qjbB-0 6.000000e-91 3 233 88.000000 see alignment COMPLEX (SIGNAL TRANSDUCTION/PEPTIDE) 23-JUN-99 1QJB CROSS_PDB NPSA
pdb2btpA-0 1.000000e-90 2 232 83.000000 see alignment COMPLEX (SIGNAL TRANSDUCTION/PEPTIDE) 05-JUN-05 2BTP CROSS_PDB NPSA
pdb2c1jA-0 2.000000e-90 3 232 88.000000 see alignment SIGNALING PROTEIN/COMPLEX 13-SEP-05 2C1J CROSS_PDB NPSA
pdb2c1jB-0 2.000000e-90 3 232 88.000000 see alignment SIGNALING PROTEIN/COMPLEX 13-SEP-05 2C1J CROSS_PDB NPSA
pdb2o02B-0 2.000000e-90 3 232 88.000000 see alignment PROTEIN BINDING/TOXIN 27-NOV-06 2O02 CROSS_PDB NPSA
pdb2c23A-0 5.000000e-90 3 232 97.000000 see alignment SIGNALING PROTEIN/COMPLEX 24-AUG-05 2C23 CROSS_PDB NPSA

pdb2c74B-0	6.000000e-84	4	232	76.000000	see alignment	SIGNALING PROTEIN/PEPTIDE COMPLEX 17-NOV-05 2C74 CROSS_PDB	NPSA
pdb2b05F-0	7.000000e-84	4	231	76.000000	see alignment	CELL CYCLE 13-SEP-05 2B05 CROSS_PDB	NPSA
pdb2b05E-0	2.000000e-83	5	231	76.000000	see alignment	CELL CYCLE 13-SEP-05 2B05 CROSS_PDB	NPSA
pdb2c63A-0	2.000000e-83	4	232	76.000000	see alignment	SIGNALING PROTEIN/PEPTIDE COMPLEX 07-NOV-05 2C63 CROSS_PDB	NPSA
pdb2c63B-0	2.000000e-83	4	232	76.000000	see alignment	SIGNALING PROTEIN/PEPTIDE COMPLEX 07-NOV-05 2C63 CROSS_PDB	NPSA
pdb2c63C-0	2.000000e-83	4	232	76.000000	see alignment	SIGNALING PROTEIN/PEPTIDE COMPLEX 07-NOV-05 2C63 CROSS_PDB	NPSA
pdb2c63D-0	2.000000e-83	4	232	76.000000	see alignment	SIGNALING PROTEIN/PEPTIDE COMPLEX 07-NOV-05 2C63 CROSS_PDB	NPSA
pdb1yz5B-0	2.000000e-82	4	232	68.000000	see alignment	SIGNALING PROTEIN 28-FEB-05 1YZ5 CROSS_PDB	NPSA
pdb1ywtB-0	4.000000e-82	3	231	67.000000	see alignment	SIGNALING PROTEIN/DE NOVO PROTEIN 18-FEB-05 1YWT CROSS_PDB	NPSA
pdb1ywtA-0	6.000000e-82	3	231	67.000000	see alignment	SIGNALING PROTEIN/DE NOVO PROTEIN 18-FEB-05 1YWT CROSS_PDB	NPSA
pdb1yz5A-0	2.000000e-81	3	233	66.000000	see alignment	SIGNALING PROTEIN 28-FEB-05 1YZ5 CROSS_PDB	NPSA
pdb2npmB-0	3.000000e-81	5	232	62.000000	see alignment	PROTEIN BINDING 27-OCT-06 2NPM CROSS_PDB	NPSA
pdb2npmA-0	7.000000e-81	5	232	62.000000	see alignment	PROTEIN BINDING 27-OCT-06 2NPM CROSS_PDB	NPSA
pdb1a37A-0	7.000000e-73	3	230	75.000000	see alignment	COMPLEX (SIGNAL TRANSDUCTION/PEPTIDE) 28-JAN-98 1A37 CROSS_PDB	NPSA
pdb1a37B-0	7.000000e-73	3	230	75.000000	see alignment	COMPLEX (SIGNAL TRANSDUCTION/PEPTIDE) 28-JAN-98 1A37 CROSS_PDB	NPSA
pdb1a4oA-0	1.000000e-72	3	230	76.000000	see alignment	SIGNAL TRANSDUCTION 01-FEB-98 1A4O CROSS_PDB	NPSA
pdb1a4oB-0	1.000000e-72	3	230	76.000000	see alignment	SIGNAL TRANSDUCTION 01-FEB-98 1A4O CROSS_PDB	NPSA
pdb1a4oC-0	1.000000e-72	3	230	76.000000	see alignment	SIGNAL TRANSDUCTION 01-FEB-98 1A4O CROSS_PDB	NPSA
pdb1a4oD-0	1.000000e-72	3	230	76.000000	see alignment	SIGNAL TRANSDUCTION 01-FEB-98 1A4O CROSS_PDB	NPSA
pdb2ijpB-0	8.000000e-40	6	234	24.000000	see alignment	SIGNALING PROTEIN 30-SEP-06 2IJP CROSS_PDB	NPSA
pdb2ijpA-0	4.000000e-35	20	236	23.000000	see alignment	SIGNALING PROTEIN 30-SEP-06 2IJP CROSS_PDB	NPSA
pdb2ijpD-0	5.000000e-35	11	229	24.000000	see alignment	SIGNALING PROTEIN 30-SEP-06 2IJP CROSS_PDB	NPSA
pdb2ijpC-0	8.000000e-35	21	236	23.000000	see alignment	SIGNALING PROTEIN 30-SEP-06 2IJP CROSS_PDB	NPSA

PSI-BLAST hasn't converged in 3 runs.

Select template



Pôle BioInformatique Lyonnais

Geno3D

Geno3D is the IBCP contribution to PBIL in Lyon, France

[HOME] [GENO3D] [HELP] [REFERENCES] [NEWS] [NPS@] [SuMo] [PBIL]

Tuesday, April 8th 2008: more options for PSI-BLAST search (see news)

Job **TEMPLATES VALIDATION ...** (ID: 19250) is running on **GENO3D** server (started on 20080529-125249).
Results will be shown below. **Please wait and don't go back.**

SECOND STEP :

Enter your e-mail address : lucagenovese10@virgilio.it

THIRD STEP :

Choose number of model to generate : 3

EXPERT OPTIONS :

Inter/intra restraints ratio : 0.9

Distance restraints cut off (Angstrum): 50.0

Margin in distance restraints (Angstrum): 0.5

Margin in angle restraints (degree) : 1.0

Maximal number of distance restraints : 20000

Save full template in superposition : ◉ No ○ Yes

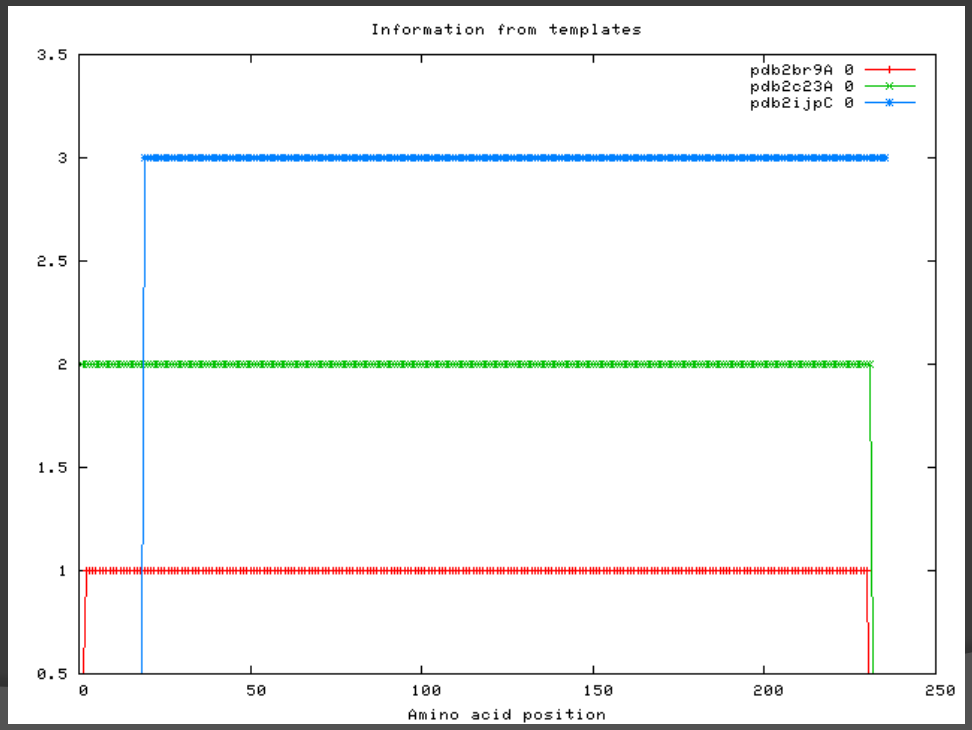
LAUNCH Geno3D

User : *public@151.81.11.42.* ***Last modification time :*** *Thu May 29 12:52:57 2008.* ***Current time :*** *Thu May 29 12:52:57 2008 This service is supported by 'Ministere de la recherche', 'Programme Bioinformatique inter-EPST', CNRS (IMABIO, COMI, GENOME) and Région Rhône-Alpes (Programme EMERGENCE) . Comments.*

-**Sequence of this chain :**
MDKSELVQKA KLAEQAERYD DMAAAMKAVT EQGHELSNEE RNLLSVAYKN VVGARRSSWR VISSIEQKTE RNEKKQQMGK EYREKIEAEL QDICNDVLEL LDKYLIPNAT QPESKVFYLK MKGDYFRYLS EVASGDNKQT TVSNSQQAYQ EAFEISKKEM QPTHPIRLGL ALNFSVFYYE ILNSPEKACS LAKTAFDEAI AELDTLNEES YKDSTLIMQL LRDNLTLWTS ENQG

Template information (Sov)	Alignment Identity	Secondary
pdb2br9A_0	68.0%	ali_antheprot ali_clustalw
pdb2c23A_0	100.0%	ali_antheprot ali_clustalw
pdb2ijpC_0	26.0%	ali_antheprot ali_clustalw

- Template at each amino acid position :



Stereochemical quality of models with PROCHECK :

Model1	core	allowed	generously	disallowed
	83.7%	14.0%	1.8%	0.5%

-ramachandran plot : jpeg postscript

-ramachandran plots for all residue types : jpeg postscript

-main-chain parameters : jpeg postscript

-side-chain parameters : jpeg postscript

-residue properties : jpeg postscript

Model 2	86.4%	11.3%	1.8%	0.5%

-ramachandran plot : jpeg postscript

-ramachandran plots for all residue types : jpeg postscript

-main-chain parameters : jpeg postscript

-side-chain parameters : jpeg postscript

-residue properties : jpeg postscript

Model 3	84.2%	14.0%	0.5%	1.4%

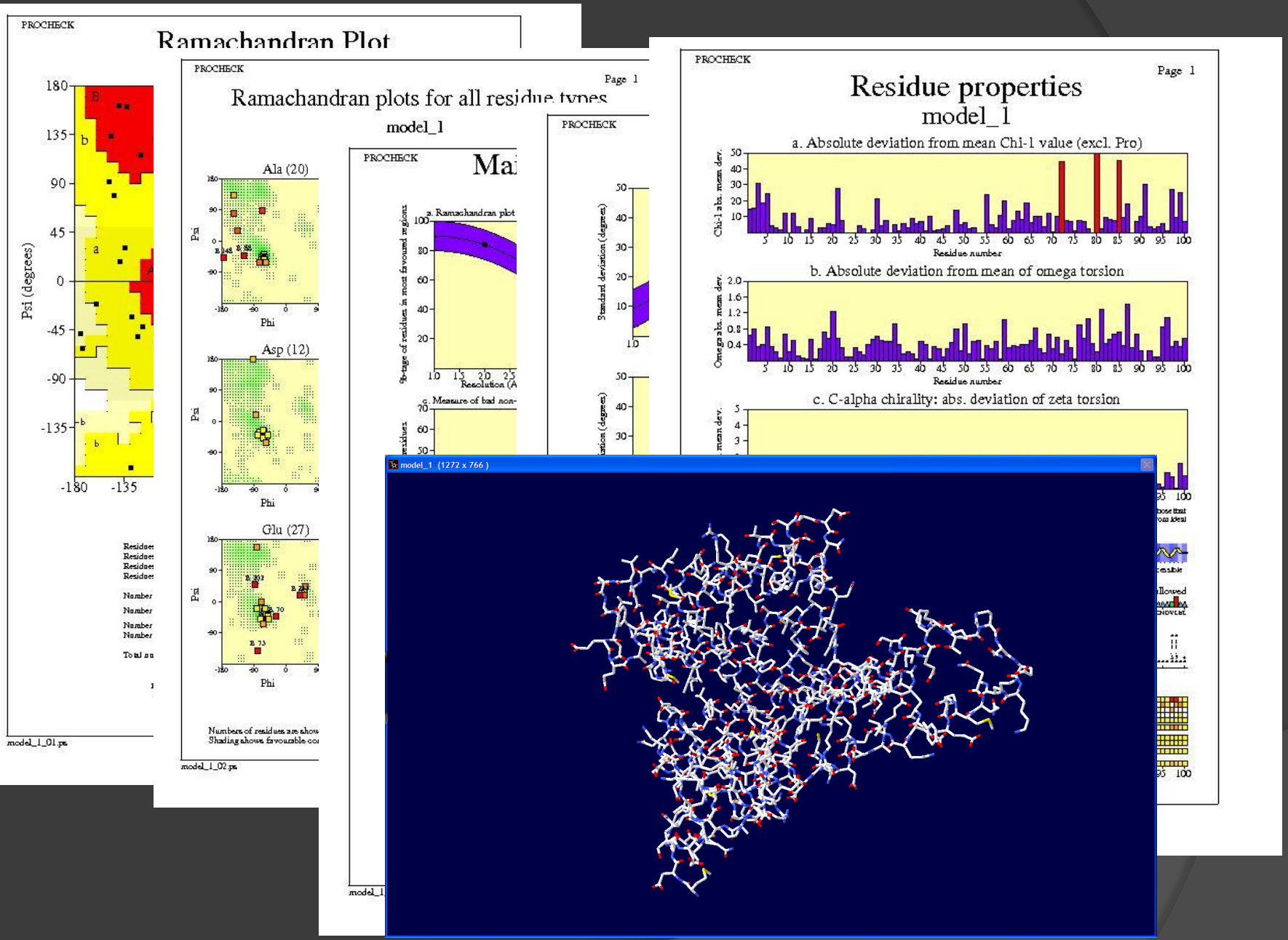
ramachandran plot : jpeg postscript

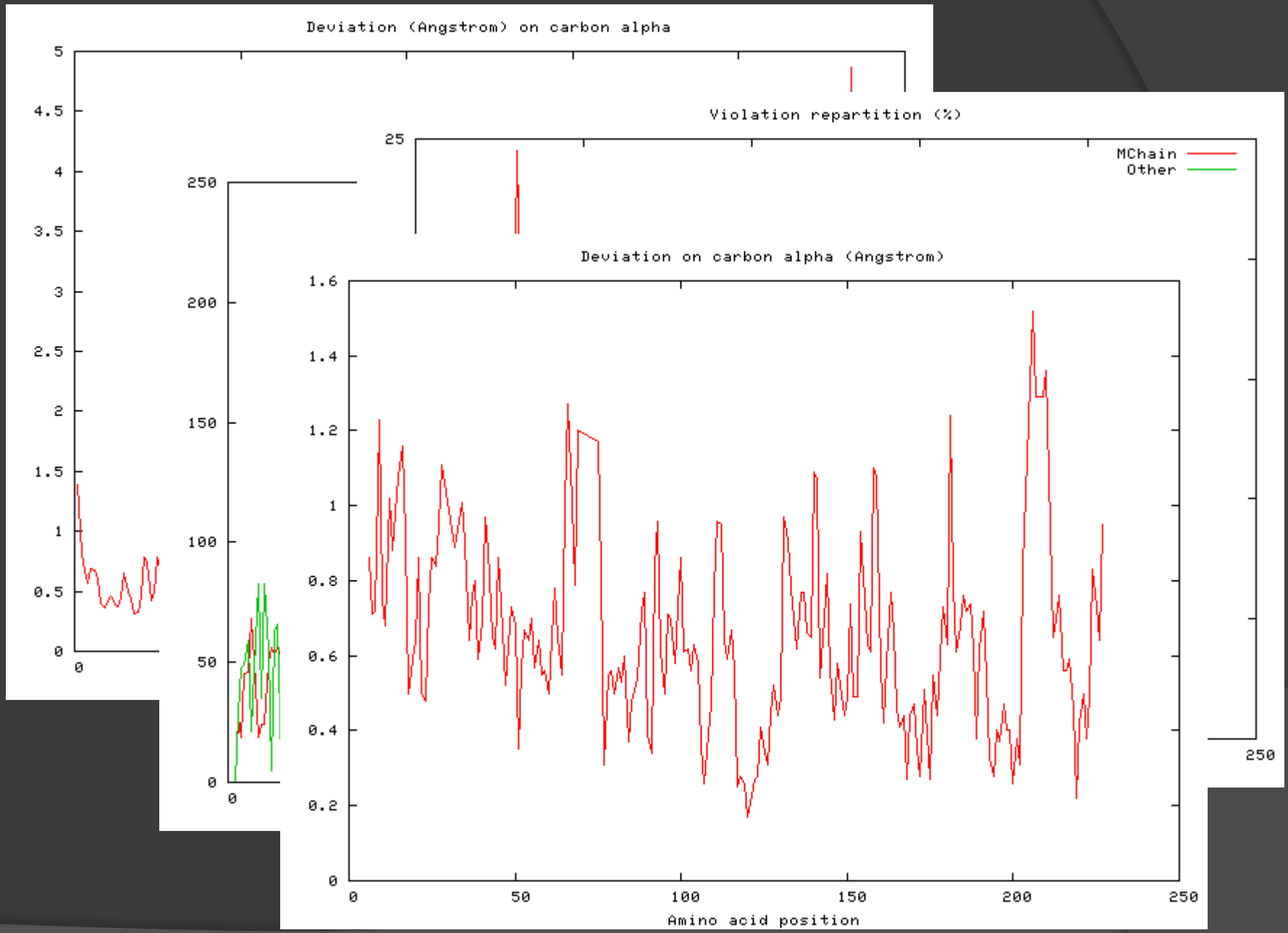
-ramachandran plots for all residue types : jpeg postscript

-main-chain parameters : jpeg postscript .

-side-chain parameters : jpeg postscript

-residue properties : jpeg postscript


PROCHECK
Ramachandran Plot
Psi (degrees)
PROCHECK
Page 1
Ramachandran plots for all residue types
model_1
Ala (20)
Asp (12)
Glu (27)
Phi
Psi
PROCHECK
a. Ramachandran plot
b. Measure of bad non-
Residue properties
model_1
a. Absolute deviation from mean Chi-1 value (excl. Pro)
b. Absolute deviation from mean of omega torsion
c. C-alpha chirality: abs. deviation of zeta torsion
Residue number
model_1 (1272 x 766)


Deviation (Angstrom) on carbon alpha
Violation repartition (%)
MChain
Other
Deviation on carbon alpha (Angstrom)
Amino acid position

-Structural agreement between models (RMSD in angstrom) :

	Model 1	Model 2	Model 3
Model 1	0.00	0.75	0.78
Model 2	0.75	0.00	0.69
Model 3	0.78	0.69	0.00

Models energy (kcal/mol):

. model 1 : -11081.70
. model 2 : -11577.00
. model 3 : -11479.00

- Number of violation of these intrachain restraints :

	MChain	Other
Model 1	205(2.70%)	111(1.40%)
Model 2	114(1.50%)	65(0.80%)
Model 3	167(2.20%)	95(1.20%)

Sviluppi futuri:

- modellizzazione di dimeri;
-Includere il ligando nel processo di modellizzazione molecolare;
-possibilità di caricare allineamenti personali.

Se ci troviamo nella situazione intermedia dobbiamo fare altre valutazioni

Confronto della predizione di struttura secondaria

Residui idrofobici interni conservati

Ponti a disolfuro conservati

Pattern funzionali conservati

Modellizzazione comparativa (o per similarità di sequenza)

Permette di costruire la struttura tridimensionale di una proteina

sulla base della SIMILARITÀ DI SEQUENZA con un'altra proteina

di struttura NOTA

che viene usata come STAMPO.

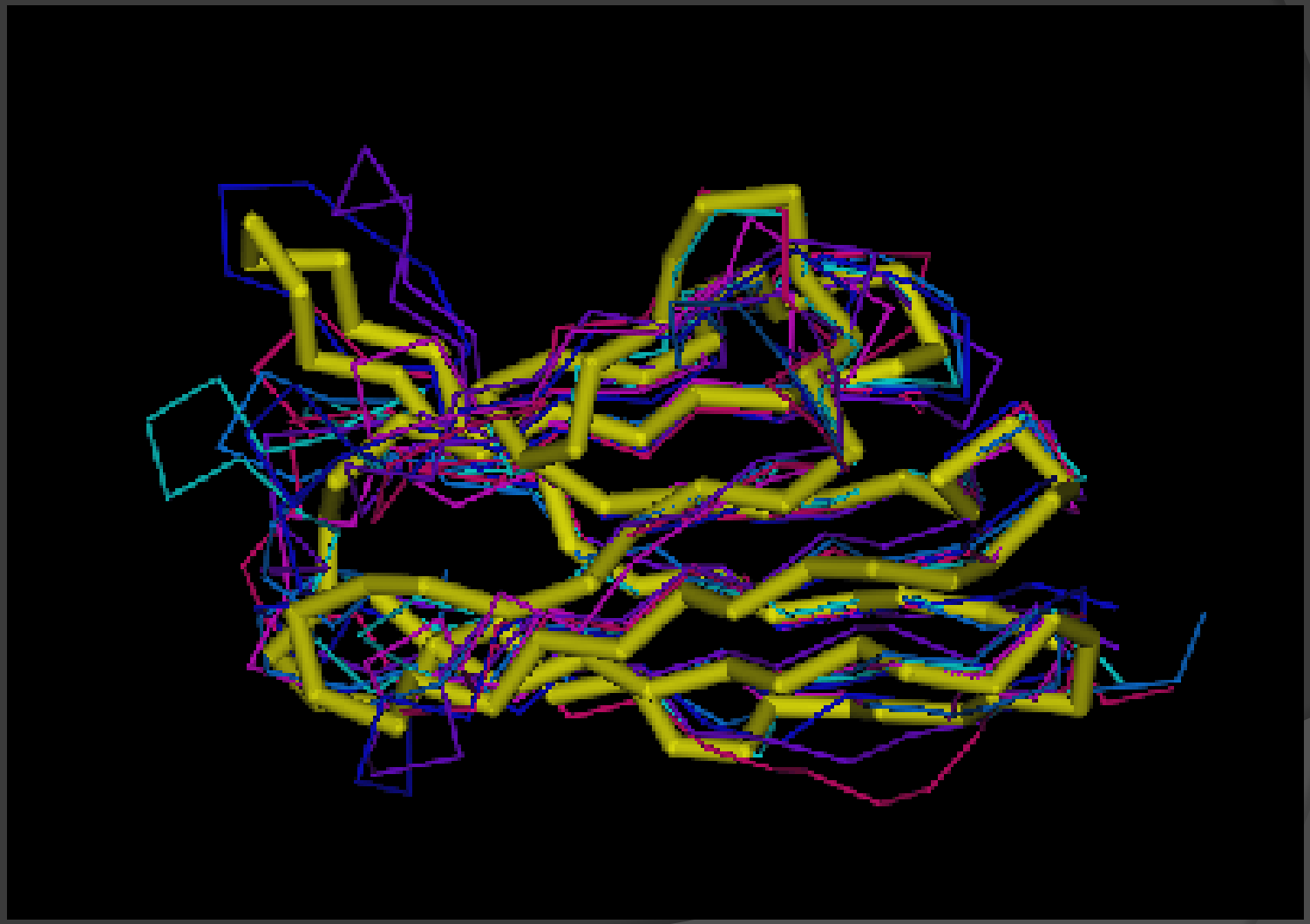
Passi fondamentali

1. Allineamento di sequenza con la/le proteina/e “stampo”

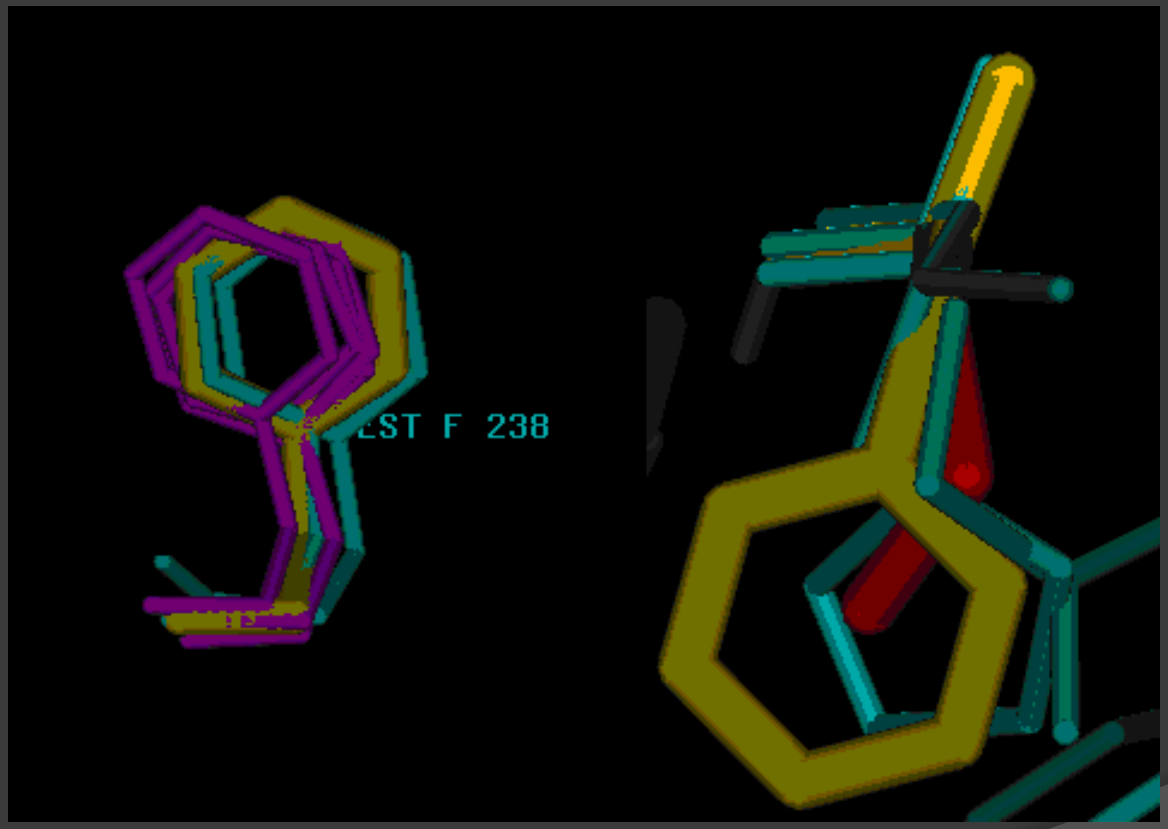
*** aa identici**
. aa simili

```
4mdh.aa     1   SEPIRVLVTG AAGQIAYSLL YSIGNGSVFG KDQPIILVLL DITPMMGVLD
11BMD       1   KAPVRVAVTG AAGQIGYSLL FRIAAGEMLG KDQPVILQLL EIPQAMKALE
                  *.** *** *****.****  . *. * ..* ****.** ** .*   *  *.

4mdh.aa    51   GVLMELQDCA LPLLKDVIAT DKEEIAFKDL DVAILVGSMP RRDGMERKDL
11BMD      51   GVVMELEDCA FPLLAGLEAT DDPDVAFKDA DYALLVGAAP RKAGMERRDL
                **.***.*** .***   . ** *  ..****  * *.***. * *. ****.**

4mdh.aa   101   LKANVKIFKC QGAALDKYAK KSVKVIVVGN PANTNCLTAS KSAPSIPKEN
11BMD     101   LQVNGKIFTE QGRALAEVAK KDVKVLVVGN PANTNALIAY KNAPGLNPRN
                *  * ***   ** **   ** * ***.**** ***** * *  *.**..   *

4mdh.aa   151   FSCLTRLDHN RAKAQIALKL GVTSDDVKNV IIWGNHSSTQ YPDVNHAKVK
11BMD     151   FTAMTRLDHN RAKAQLAKKT GTGVDRIRRM TVWGNHSSTM FPDLFHAEVD
                *. .****** *****.* *  *   * .. .  .*******  .**. ** *

4mdh.aa   201   LQAKEVGVYE AVKDDSWLKG EFITTVQQRG AAVIKARKLS SAMSAAKAIC
11BMD     201   GKP----AIE LVDME-WYEK VFIPTVAQRG AAIIQARGAS SAASAANAAI
                 ..       *  *  .      ** ** *** **.* **  * ** ***.*

4mdh.aa   251   DHVRDIWFGT PEGEFVSMGI ISDGNSYGVP DDLLYSFPVT IKDKTWKIVE
11BMD     246   EHIRDWALGT PEGDWVSMAV PSQGE-YGIP EGIVYSFPVT AKDGAYRVVE
                .*.**  .** ***. ***..  *.*. **.* . ..******  ** . ..**

4mdh.aa   301   GLPINDFSRE KMDLTAKELA EEKETAFEFL SSA
11BMD     295   GLEINEFARK RMEITAQELL DEMEQVKALG LI
                ** **.*.*  .*..** **  .*.*    .
```

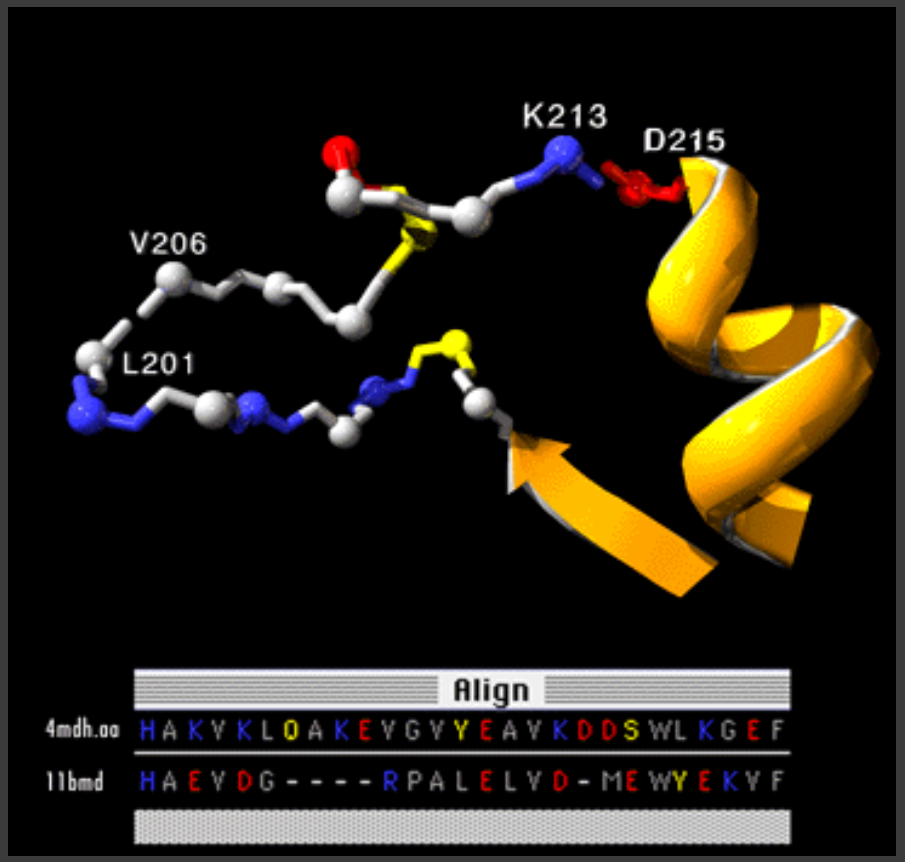
2. Costruzione dello scheletro

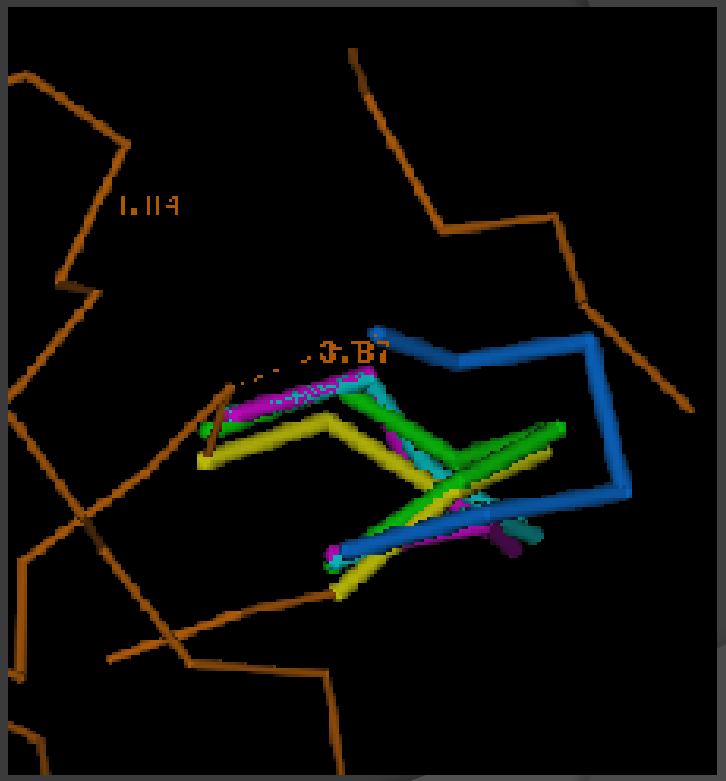


3. Visualizzazione delle catene laterali e risoluzione di ingombri sterici



4. Inserimento dei loop corrispondenti a “buchi” nell’allineamento





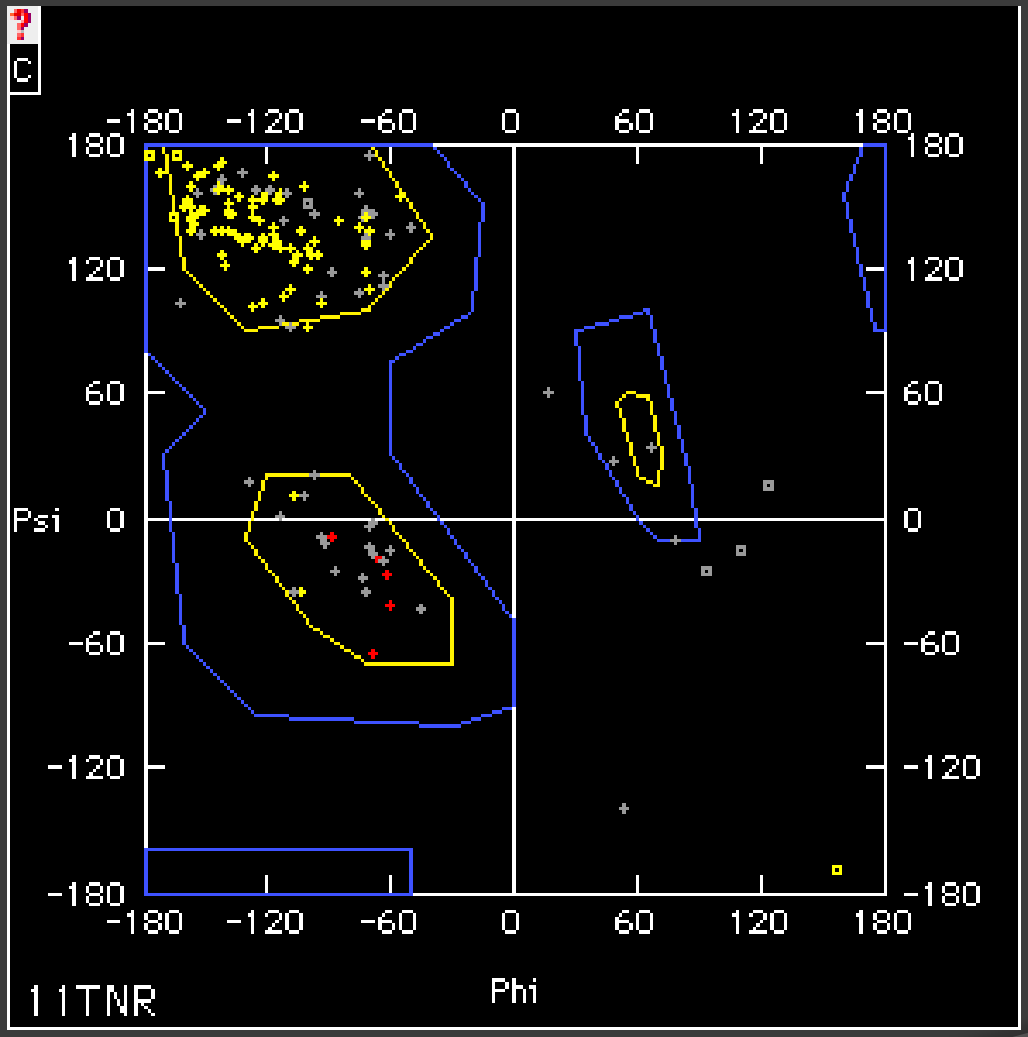
5. Ottimizzazione del modello

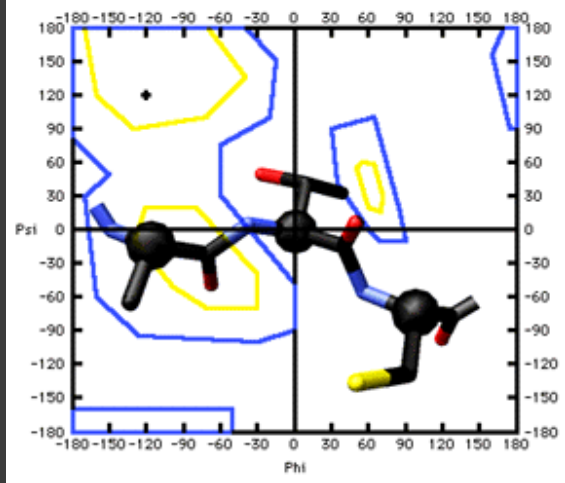
Regolarizzazione di legami, angoli e torsioni

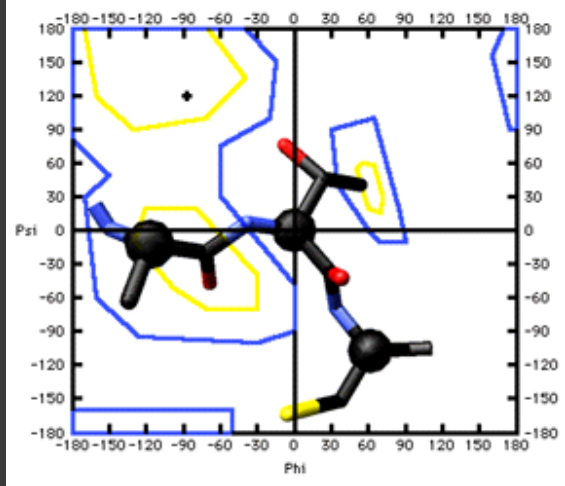
Eliminazioni di clash strutturali

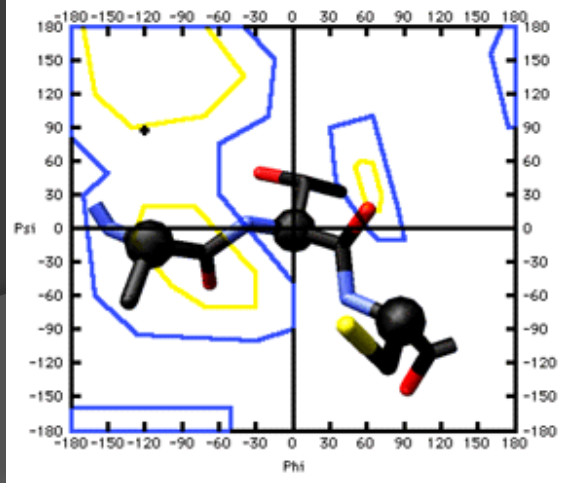
Minimizzazione energetica

6. Controllo della qualità del modello









Le proteine cambiano molto rapidamente la loro sequenza.

Difficile indicare il grado di similarità necessario per dimostrare in modo non ambiguo che due proteine siano OMOLOGHE.

Doolittle dichiara:

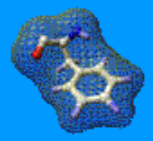
1. Se due sequenze sono più lunghe di 100 aa e l'identità è maggiore del 25% esse sono plausibilmente correlate

2. Se l'identità è tra il 15-25 % potrebbero essere correlate

3. Se l'identità è sotto il 15% probabilmente non sono correlate

ExPASy Home page | Site Map | Search ExPASy | Contact us

Search Swiss-Prot/TrEMBL for hemagglutinin influenz Go Clear



- Download
- User Guide
- Tips & Tricks
- Tutorial
- Feedback
- Index
- Art Gallery
- Mirror sites
- References

This mirror site is hosted by

GlaxoSmithKline R&D
&
the Swiss Institute of Bioinformatics

present

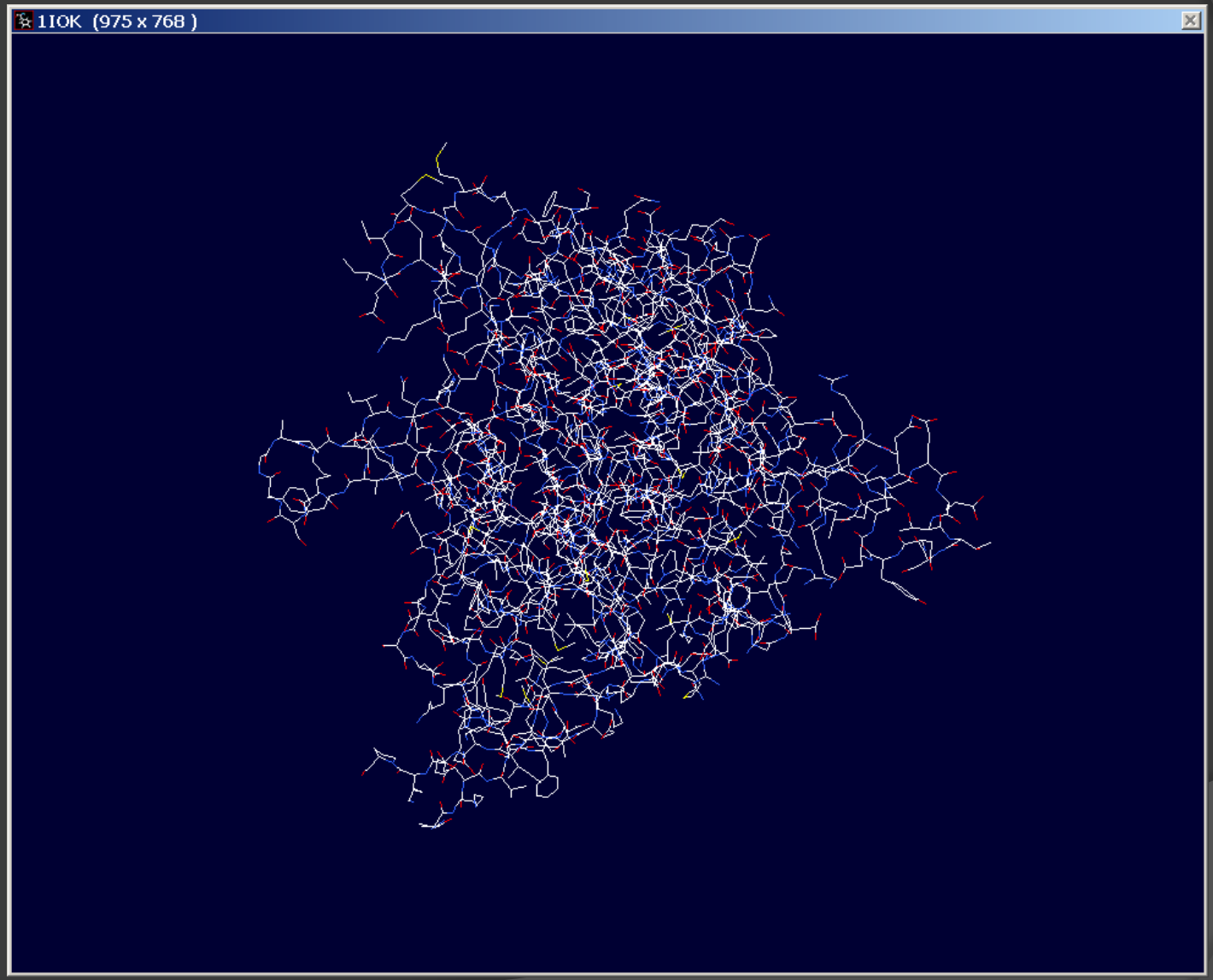
Deep View
Swiss-PdbViewer

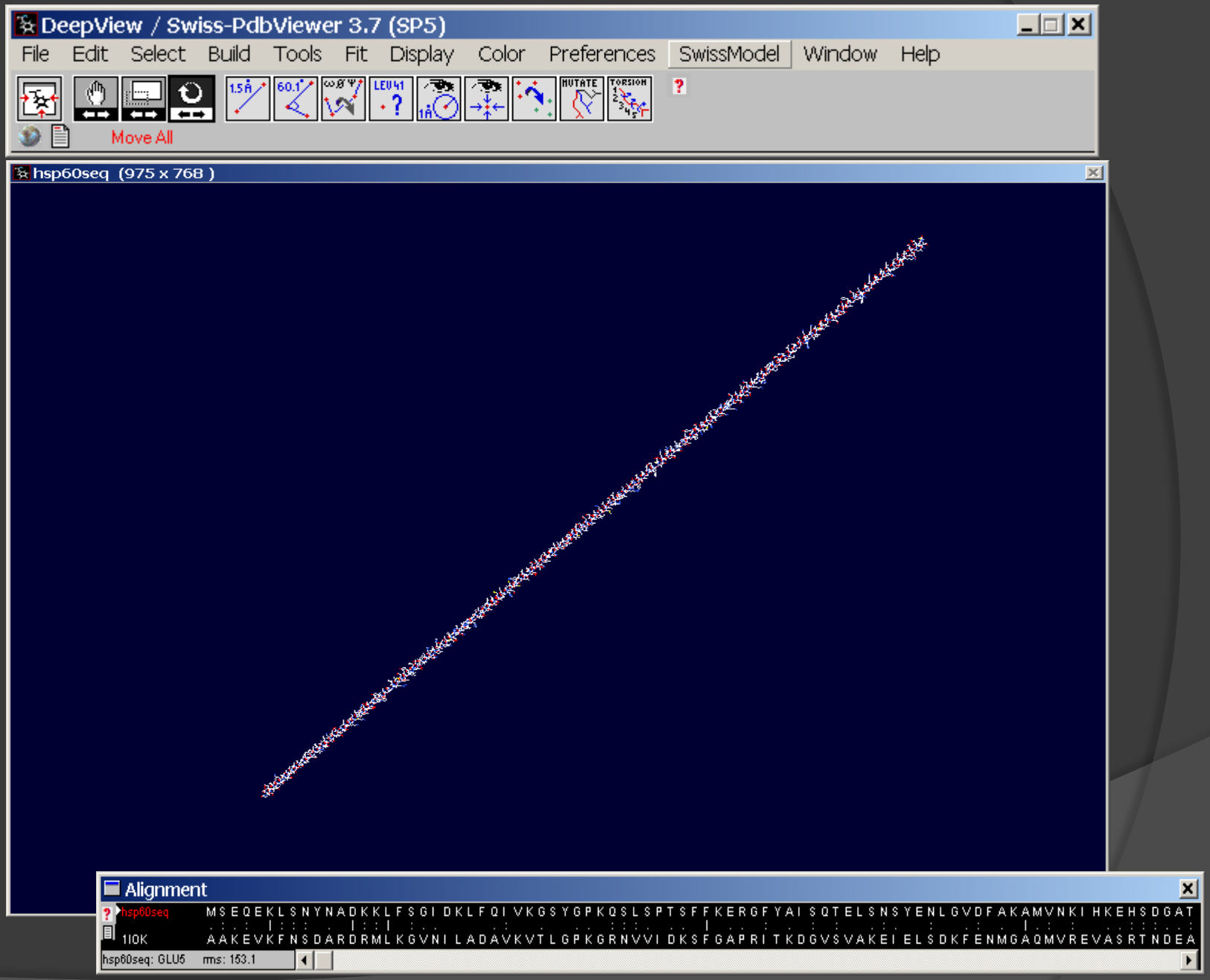
by

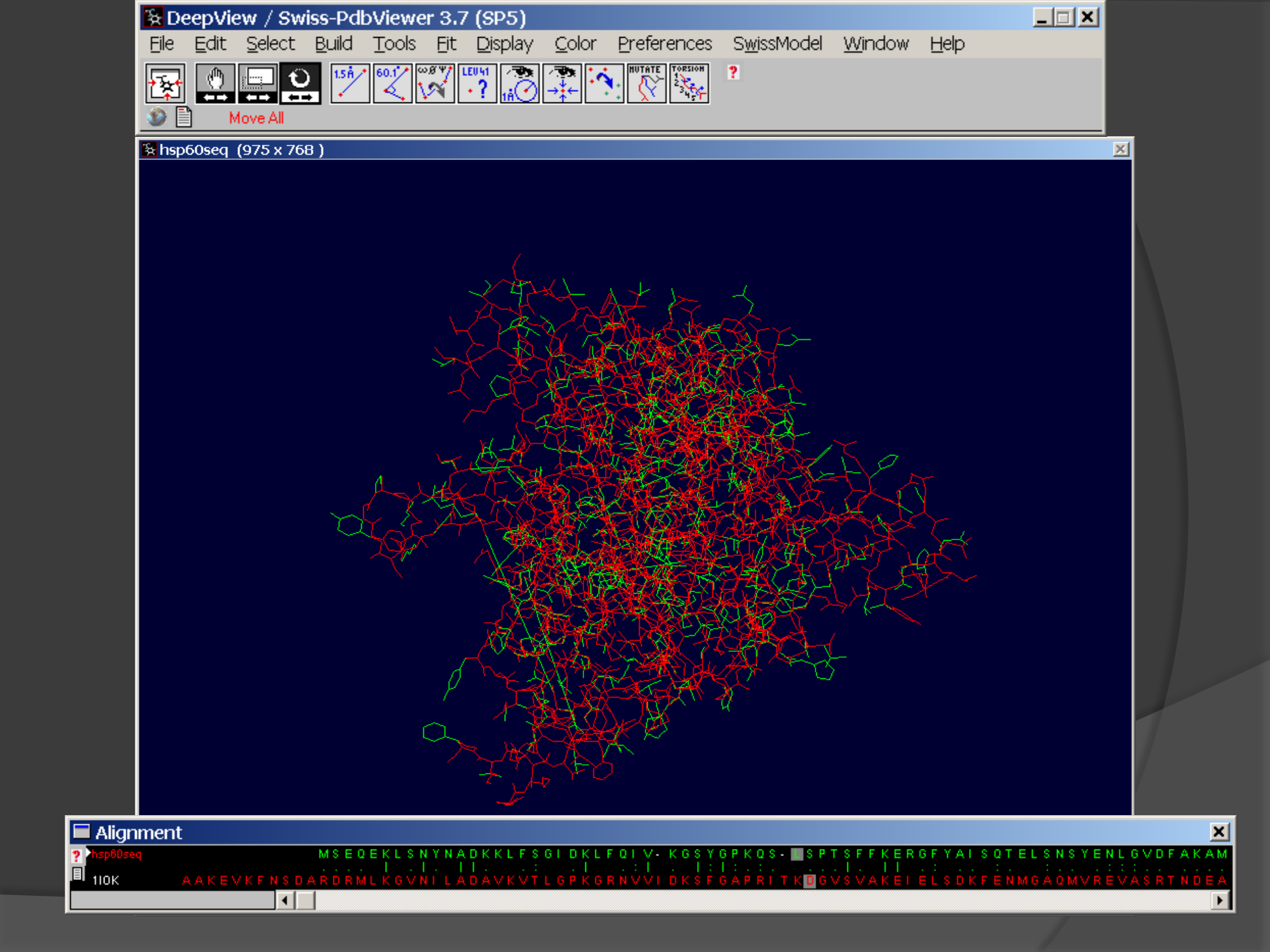
Nicolas Guex , *Alexandre Diemand* , *Manuel C. Peitsch* , & *Torsten Schwede*

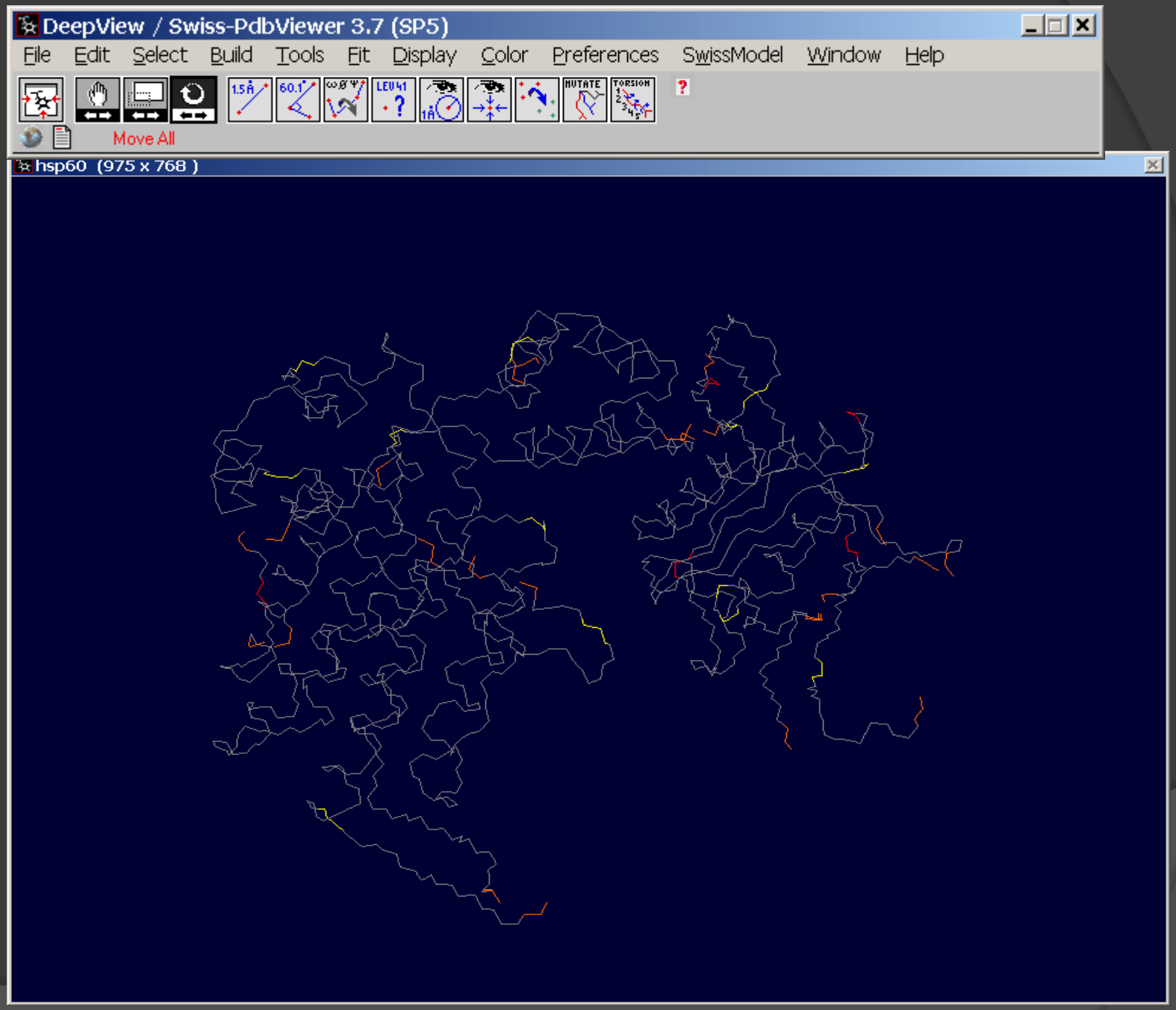
News:

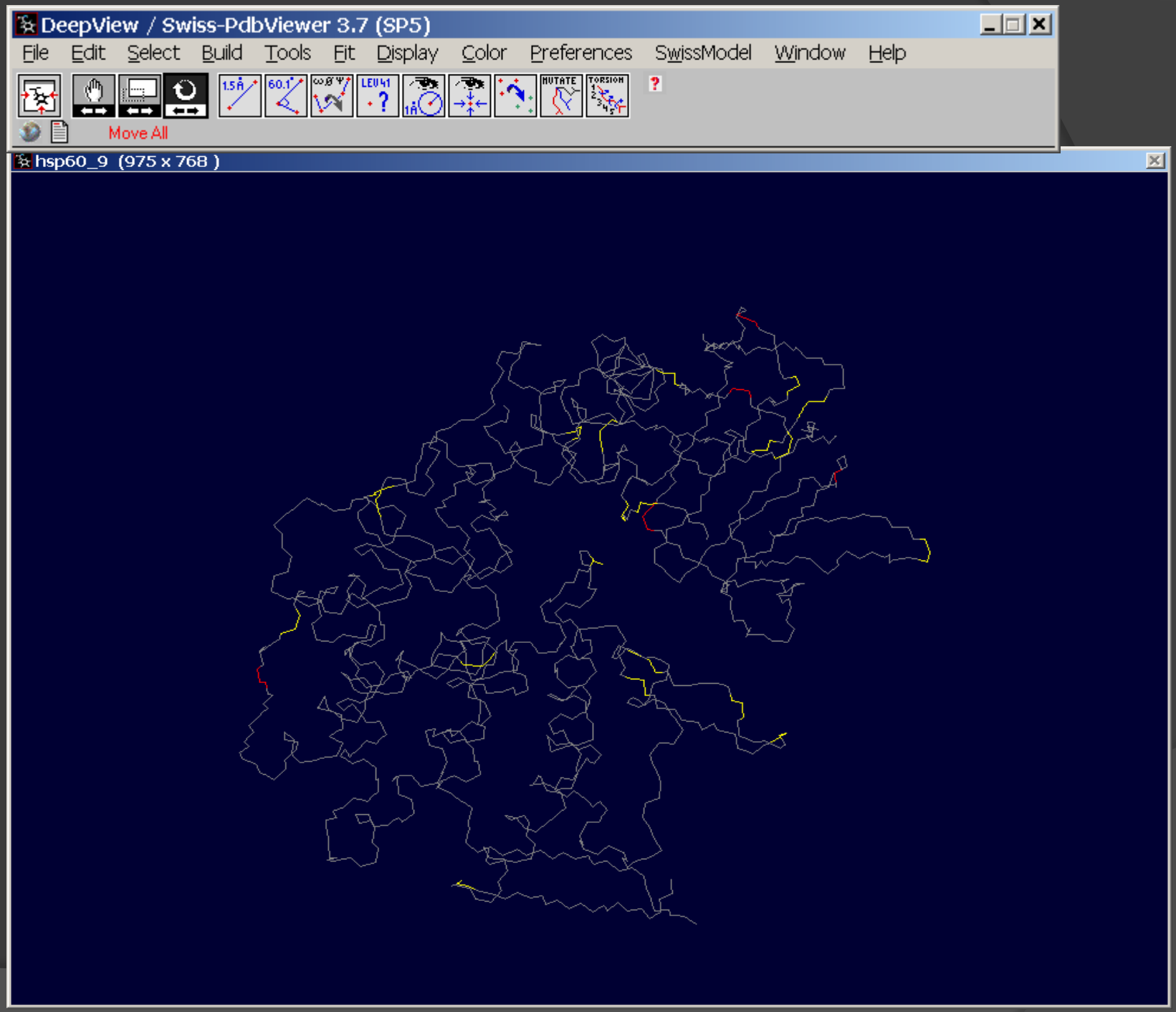
- Please update your preferences for the DeepView Network Service
- Current version: Deep View - spdbv 3.7

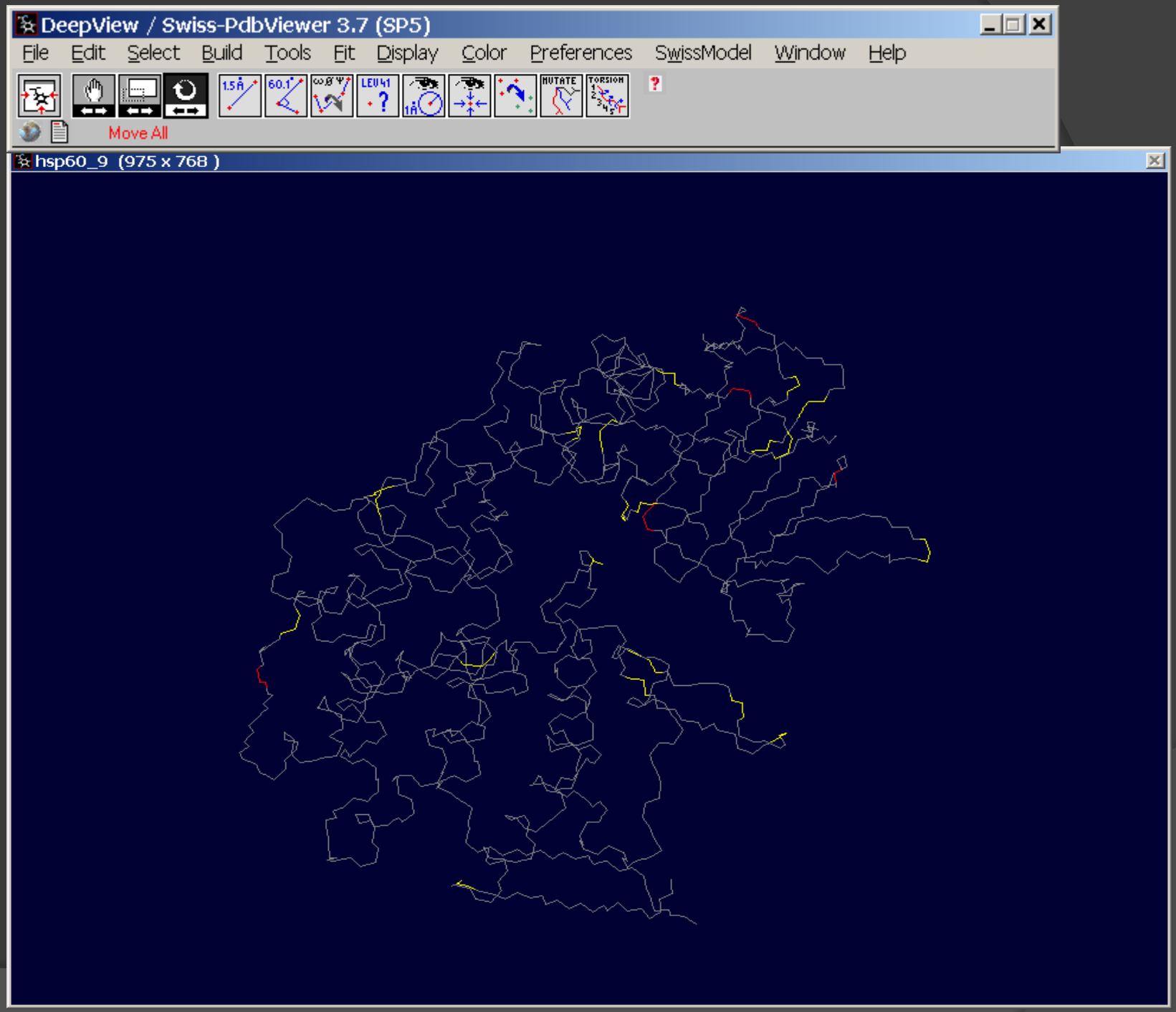

1IOK (975 x 768)

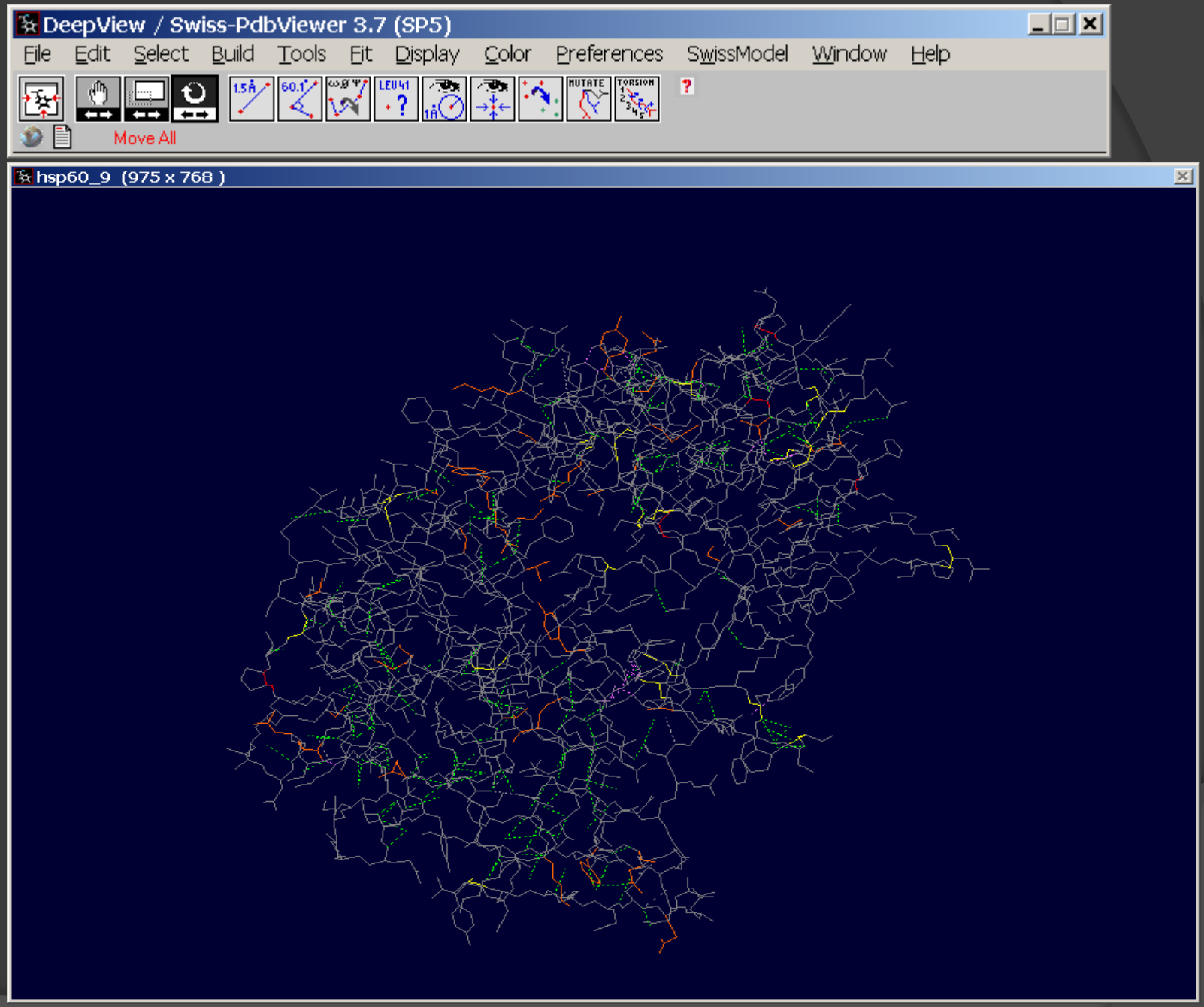

DeepView / Swiss-PdbViewer 3.7 (SP5)
File Edit Select Build Tools Fit Display Color Preferences SwissModel Window Help
Move All
hsp60seq (975 x 768)
Alignment
hsp60seq MSEQEKLSNYNADKKLFSGIDKLFQIVKGSYGPKQSLSPTSFFKERGFYAISQTELSNSYENLGVDFAKAMVNKIHKEHSDGAT
1IOK AAKEVKFNSDARDRMLKGVNILADAVKVTLGPKGRNVVIDKSFGAPRITKDGVSVAKEIELSDKFENMGAQMVREVASRTNDEA
hsp60seq: GLU5 rms: 153.1


DeepView / Swiss-PdbViewer 3.7 (SP5)
File Edit Select Build Tools Fit Display Color Preferences SwissModel Window Help
Move All
hsp60seq (975 x 768)
Alignment
hsp60seq
MSEQEKLSNYNADKKLFSGIDKLFQIV-KGSYGPKQS-SPTSFFKERGFYAISQTELSNSYENLGVDFAKAM
110K
AAKEVKFNSDARDRMLKGVNILADAVKVTLGPKGRNVVIDKSFGAPRITKDGVSVAKEIELSDKFENMGAQMVREVASRTNDEA


DeepView / Swiss-PdbViewer 3.7 (SP5)
File Edit Select Build Tools Fit Display Color Preferences SwissModel Window Help
Move All
hsp60 (975 x 768)


DeepView / Swiss-PdbViewer 3.7 (SP5)
File Edit Select Build Tools Fit Display Color Preferences SwissModel Window Help
Move All
hsp60_9 (975 x 768)


DeepView / Swiss-PdbViewer 3.7 (SP5)
File Edit Select Build Tools Fit Display Color Preferences SwissModel Window Help
Move All
hsp60_9 (975 x 768)


DeepView / Swiss-PdbViewer 3.7 (SP5)
File Edit Select Build Tools Fit Display Color Preferences SwissModel Window Help
Move All
hsp60_9 (975 x 768)

M:\local\wbin\deepview37sp5\temp\hsp60_9.E1

*LEU	438	0.957	1.778	4.831	2.419	1412.27	-0.27	0.0000 // E=	1421.982
LYSH	439	0.356	2.717	1.912	1.177	-26.96	-6.46	0.0000 // E=	-27.259
LEU	440	1.311	0.734	3.539	0.240	547.80	-6.51	0.0000 // E=	547.117
*LEU	441	1.022	1.911	1.916	0.006	1400.84	-1.57	0.0000 // E=	1404.128
ALA	442	1.459	0.579	2.317	0.621	-26.89	-13.46	0.0000 // E=	-35.374
THR	443	0.360	0.979	2.199	0.596	5.62	-16.20	0.0000 // E=	-6.444
ASN	444	1.969	0.979	3.857	0.500	158.27	-155.41	0.0000 // E=	10.162
ALA	445	0.941	1.571	2.969	0.040	125.52	6.62	0.0000 // E=	137.669
ASP	446	1.253	2.531	5.586	1.638	103.25	-6.67	0.0000 // E=	107.587
LEU	447	0.731	1.262	3.441	0.461	20.67	1.45	0.0000 // E=	28.013
ASP	448	0.428	2.998	1.492	0.019	-4.60	15.61	0.0000 // E=	15.943
GLY	449	0.651	4.519	1.317	0.221	-10.35	24.68	0.0000 // E=	21.037
ASP	450	1.380	0.482	3.657	1.642	-17.89	11.91	0.0000 // E=	1.173
ALA	451	0.962	1.162	3.081	0.073	-8.43	-3.29	0.0000 // E=	-6.441
VAL	452	2.441	3.877	1.873	1.924	5.50	-0.24	0.0000 // E=	15.381
ILE	453	0.161	4.876	3.165	1.495	-17.45	-3.05	0.0000 // E=	-10.801
ALA	454	0.921	1.594	0.539	1.009	-13.00	3.19	0.0000 // E=	-5.752
*LYSH	455	2.422	1.094	19.372	0.417	49888.76	1.07	0.0000 // E=	49913.137
*LEU	456	3.081	1.742	1.621	0.439	1094.58	0.94	0.0000 // E=	1102.408
SER	457	7.411	25.125	4.264	1.514	19.54	-6.72	0.0000 // E=	51.130
SER	458	0.304	2.984	4.219	0.016	-5.72	-11.67	0.0000 // E=	-9.862
*LEU	459	0.844	1.922	2.514	0.011	51025.36	48.02	0.0000 // E=	51078.676
GLY	460	0.342	1.122	0.887	0.043	-5.97	54.92	0.0000 // E=	51.349
THR	461	0.289	2.615	3.382	1.272	-11.29	-1.08	0.0000 // E=	-4.811
THR	462	0.750	5.123	3.009	1.359	-8.95	0.60	0.0000 // E=	1.884
SER	463	0.536	1.606	2.512	0.493	-17.45	-10.70	0.0000 // E=	-23.005
LEU	464	1.823	6.804	3.995	1.526	29.20	34.46	0.0000 // E=	77.806
GLY	465	1.349	1.478	2.197	2.839	33.23	40.19	0.0000 // E=	81.283
ILE	466	0.835	5.533	16.903	1.263	-14.95	-0.43	0.0000 // E=	9.157
SER	467	0.338	0.808	2.261	1.532	49.52	-26.64	0.0000 // E=	27.826
VAL	468	0.415	3.748	2.012	0.042	63.38	10.86	0.0000 // E=	80.458
PHE	469	2.341	20.725	1.876	0.374	-17.44	24.90	0.0000 // E=	32.772
SER	470	0.565	1.405	2.308	0.015	-3.16	16.55	0.0000 // E=	17.684
ARG	471	1.133	5.067	2.547	1.048	23.79	-248.34	0.0000 // E=	-214.759
GLU	472	1.020	2.174	3.459	0.858	-18.37	-3.90	0.0000 // E=	-14.761
*ILE	473	0.459	1.286	3.433	0.517	2139.81	-5.65	0.0000 // E=	2139.862
*GLU	474	2.499	3.981	3.249	2.433	2095.46	-3.45	0.0000 // E=	2104.176
ASP	475	0.842	2.193	5.190	0.003	-10.37	-8.98	0.0000 // E=	-11.127
LEU	476	0.612	2.546	2.252	0.023	111.61	-18.97	0.0000 // E=	98.074
ILE	477	0.732	1.507	3.281	0.564	25.14	4.73	0.0000 // E=	35.948
ALA	478	0.232	0.449	0.031	0.590	-8.04	45.89	0.0000 // E=	39.148
GLY	479	0.318	2.249	0.275	0.003	-5.53	77.83	0.0000 // E=	75.142
GLY	480	0.945	0.235	2.916	1.280	-1.03	44.73	0.0000 // E=	49.071
ILE	481	0.743	2.333	2.782	0.019	63.08	-8.55	0.0000 // E=	60.411
LEU	482	0.998	9.213	4.624	0.872	-12.72	-5.35	0.0000 // E=	-2.366
ASP	483	0.338	1.031	5.407	0.438	-31.30	-10.59	0.0000 // E=	-34.670
SER	484	0.166	1.347	3.335	3.871	-26.02	-18.65	0.0000 // E=	-35.949
LEU	485	2.528	12.989	4.190	3.032	-2.93	-13.66	0.0000 // E=	6.149
ALA	486	0.709	2.149	0.817	0.864	-17.93	3.42	0.0000 // E=	-9.968
THR	487	0.570	1.239	3.199	0.921	4.24	-9.27	0.0000 // E=	0.894
THR	488	0.651	0.864	2.908	1.515	1.41	-9.32	0.0000 // E=	-1.971
SER	489	0.788	1.513	6.150	0.014	-21.80	-7.40	0.0000 // E=	-20.733
THR	490	1.795	1.121	0.913	1.445	-13.68	-27.75	0.0000 // E=	-36.153
ILE	491	0.399	0.960	2.818	0.696	345.12	-9.60	0.0000 // E=	340.397
LEU	492	0.372	4.440	1.394	0.029	31.28	-1.65	0.0000 // E=	35.870
ALA	493	0.826	0.496	2.168	1.521	24.50	-9.82	0.0000 // E=	19.687
GLN	494	0.392	2.690	4.304	0.015	14.39	-167.21	0.0000 // E=	-145.419
*ALA	495	0.844	0.323	2.039	0.851	719.68	-25.20	0.0000 // E=	698.546
LEU	496	0.894	1.218	1.866	0.934	67.96	-3.54	0.0000 // E=	69.333
ASP	497	1.417	0.563	3.898	1.211	27.04	-0.80	0.0000 // E=	33.324
THR	498	3.828	4.124	6.104	2.115	25.05	-17.02	0.0000 // E=	24.209
ALA	499	1.175	0.444	0.254	0.723	29.82	-3.80	0.0000 // E=	28.612
ILE	500	1.250	1.777	3.004	0.973	34.24	5.12	0.0000 // E=	46.360
LEU	501	1.881	3.150	2.349	0.001	-17.20	3.32	0.0000 // E=	-6.492
VAL	502	1.402	0.615	2.950	1.209	-0.26	-1.16	0.0000 // E=	4.763
LEU	503	7.358	18.588	6.675	0.649	38.65	-5.22	0.0000 // E=	66.702
SER	504	0.920	4.101	3.520	0.546	-0.22	-15.38	0.0000 // E=	-6.512
SER	505	0.208	3.488	4.181	0.018	-21.12	-0.40	0.0000 // E=	-13.628
LYSH	506	1.048	3.025	1.790	0.206	-12.00	4.45	0.0000 // E=	-1.484
ILE	507	1.261	1.977	3.291	1.900	19.03	1.87	0.0000 // E=	29.333
LEU	508	0.355	3.594	6.490	1.115	-19.48	-12.51	0.0000 // E=	-20.436
ILE	509	3.071	1.031	3.875	1.524	-2.56	1.85	0.0000 // E=	8.784
LEU	510	6.426	31.127	1.888	0.751	38.24	-13.81	0.0000 // E=	64.616
GLU	511	4.445	3.219	23.199	0.435	-24.44	-6.99	0.0000 // E=	-0.127
ASN	512	0.883	1.199	4.421	0.467	-16.04	-151.52	0.0000 // E=	-160.590
GLN	513	1.206	5.582	4.264	0.442	1.90	-74.12	0.0000 // E=	-60.727
OXT	513	0.000	0.000	0.000	0.000	-3.24	0.99	0.0000 // E=	-2.253
KJ/mol		3338.110	6302.236	2884.081	1743.990	240661.61	-7381.23	0.0000 // E=	247548.797

PROCHECK v.3.5.4

Roman A Laskowski, Malcolm W MacArthur, David K Smith, David T Jones, E Gail Hutchinson, A Louise Morris, David S Moss & Janet M Thornton

Checks the **stereochemical quality** of a protein structure, producing a number of PostScript plots analysing its **overall** and **residue-by-residue** geometry.

The plots can be in colour, if required.


How to run the program


Operating Manual


Checks carried out


Sample outputs


References

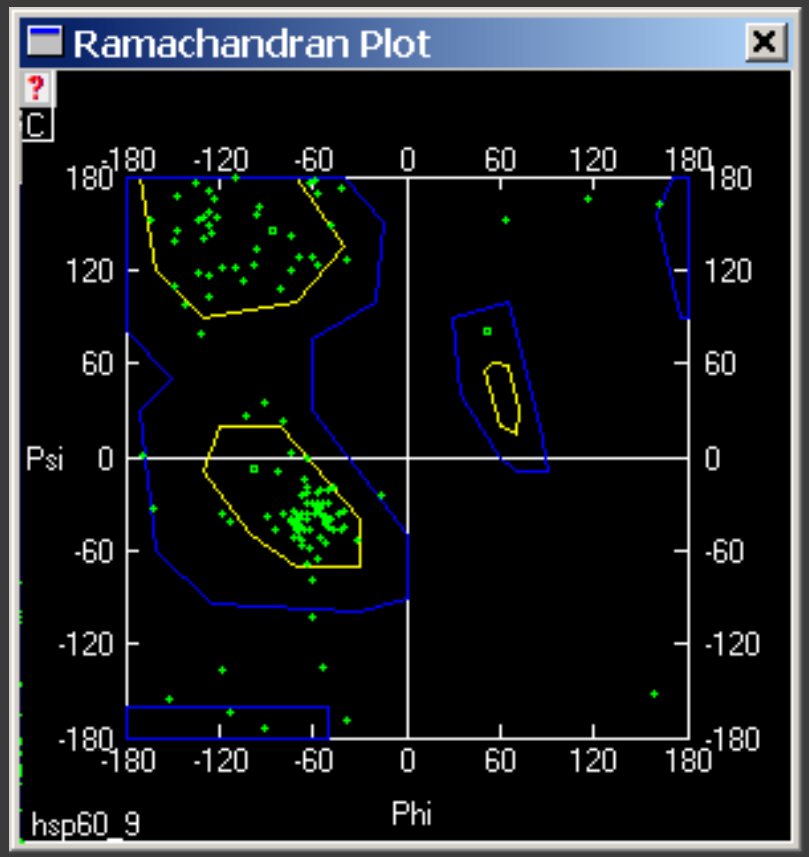
Availability

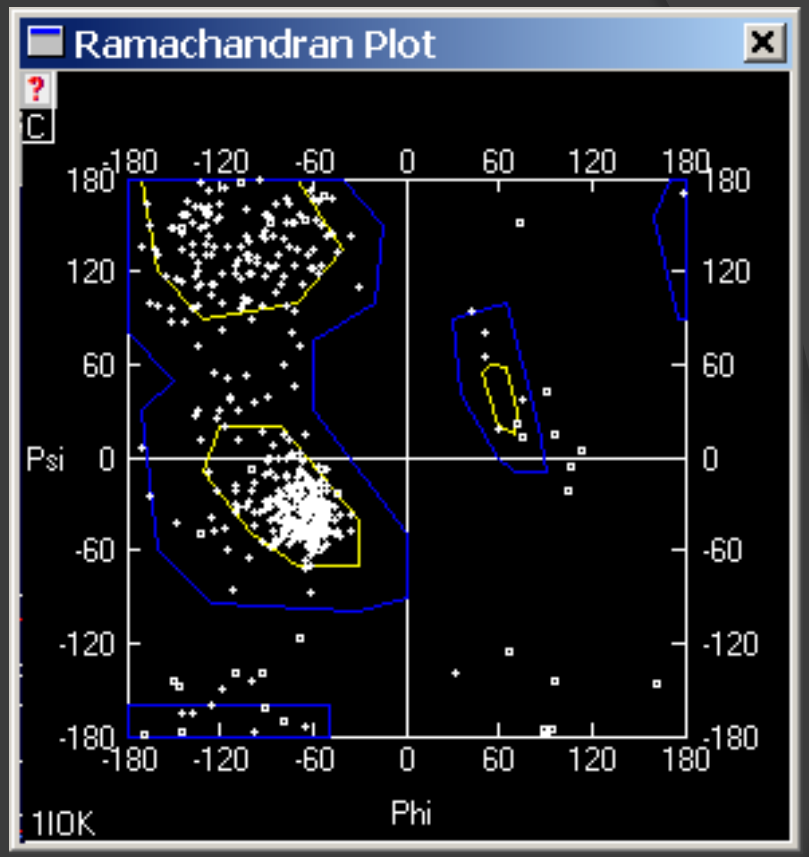
Available by anonymous ftp on: *ftp.biochem.ucl.ac.uk*

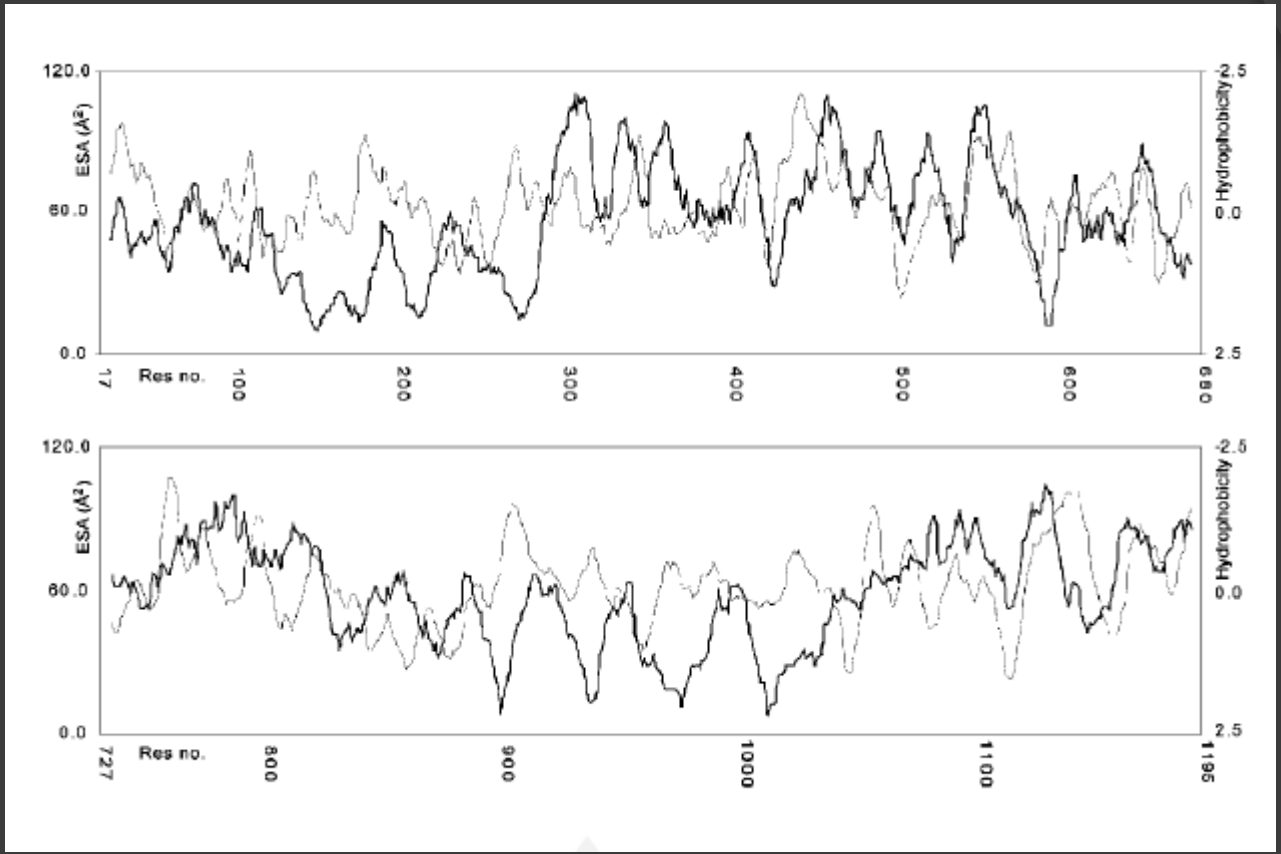
Source code can be picked up from one of two directories:-

pub/procheck/tar3_5
pub/procheck/source3_5

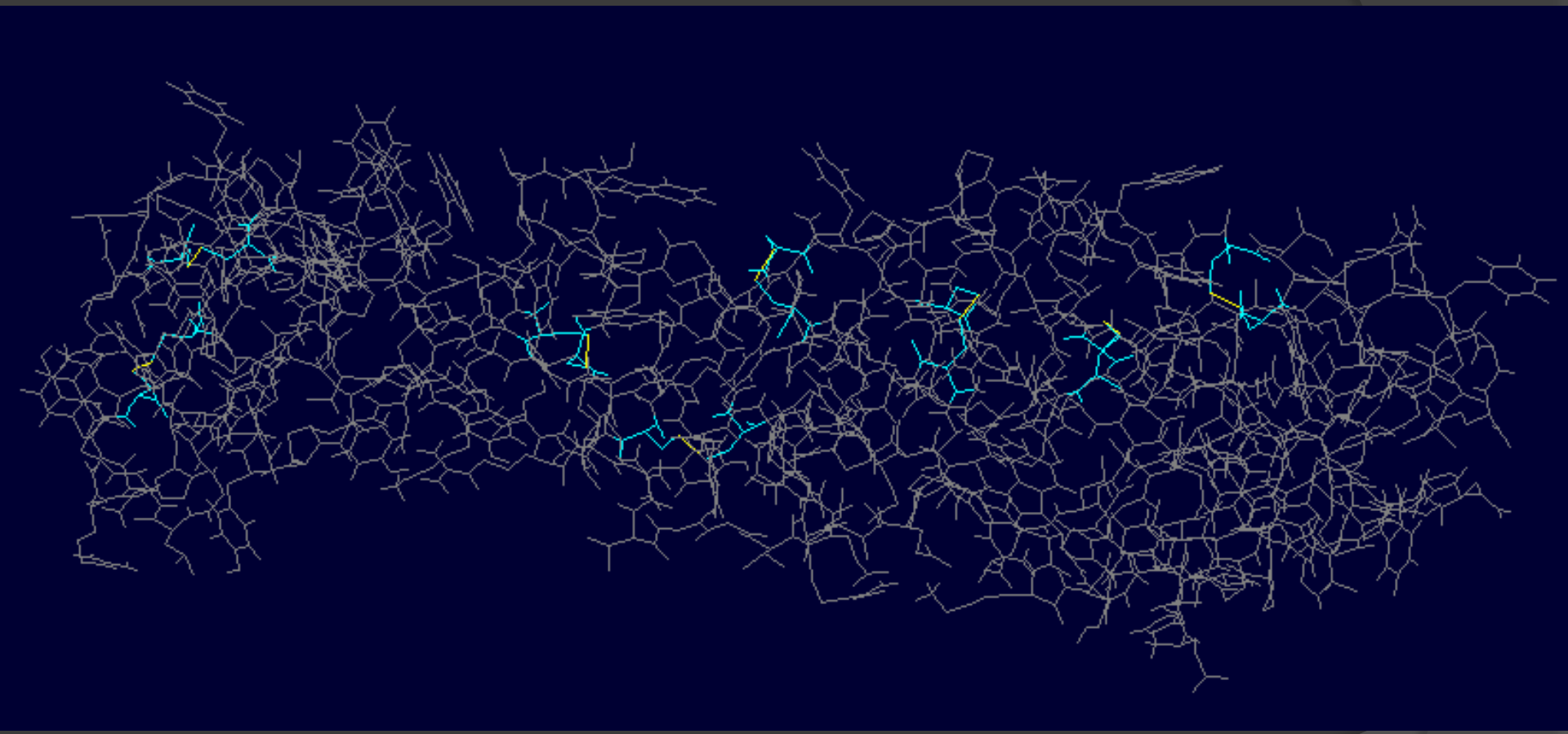
The first contains all the files concatenated into a single tar file (which is the most convenient for unix systems), and the second contains all the individual files


Ramachandran Plot
?
C
180
120
60
0
-60
-120
-180
-180
-120
-60
0
60
120
180
Psi
Phi
hsp60_9

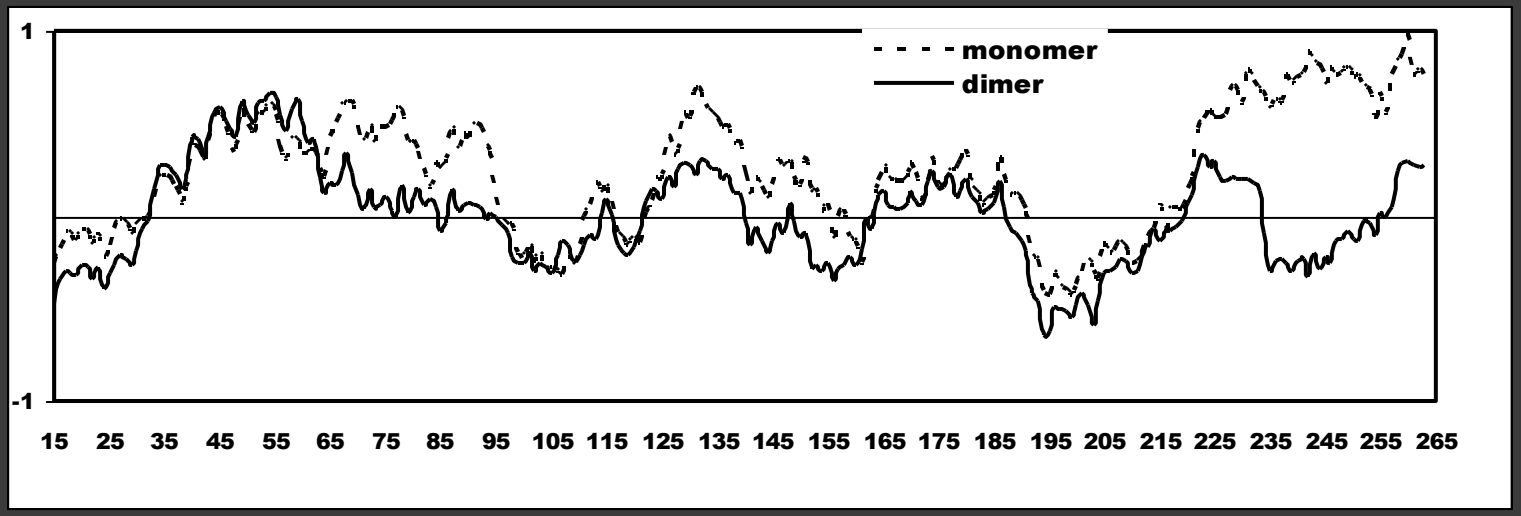

Ramachandran Plot
?
C
180
120
60
0
-60
-120
-180
-180
-120
-60
0
60
120
180
Psi
Phi
1IOK



Siti di glicosilazione, siti attivi, siti antigenici….devono essere esposti



PROSA II


1
-1
monomer
dimer
15 25 35 45 55 65 75 85 95 105 115 125 135 145 155 165 175 185 195 205 215 225 235 245 255 265

Quando l’identità di sequenza fra proteine si trova sotto il 25%

un altro metodo di procedere è quello di utilizzare metodi di riconoscimento di fold basati su **profili** e quelli detti di **THREADING**

Metodi basati sui profili:

Ciascun aa ha delle proprietà che possono essere derivate dall'analisi di proteine a struttura nota per ogni aa

Frequenza relativa ad essere presente nelle strutture secondarie note

Frequenza con cui è osservato sulla superficie di una proteina

Frequenza con cui è osservato in un ambiente idrofobico

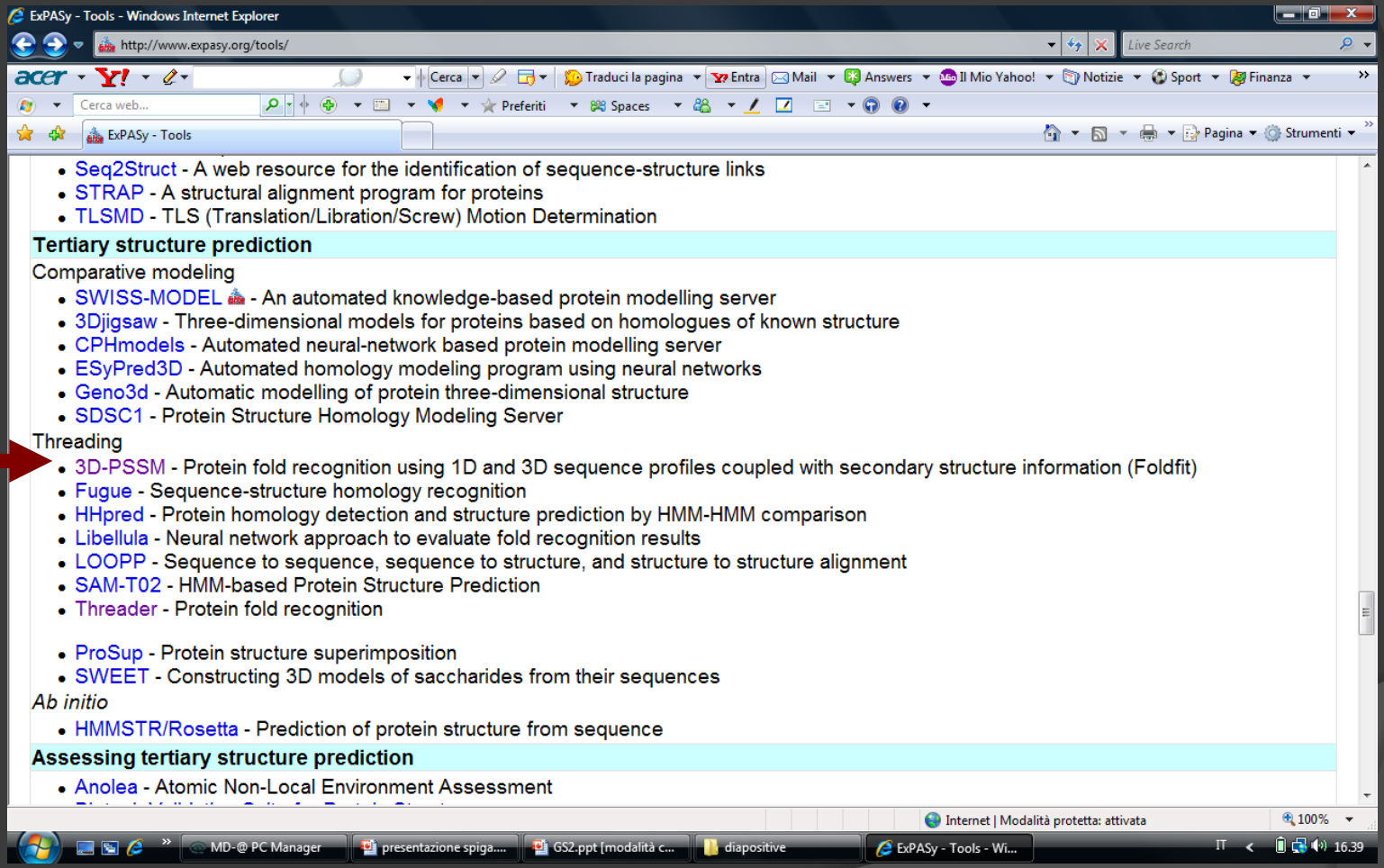
Ci sono 18 combinazioni 3 per struttura secondaria 3 per l'esposizione 2 per l'ambiente

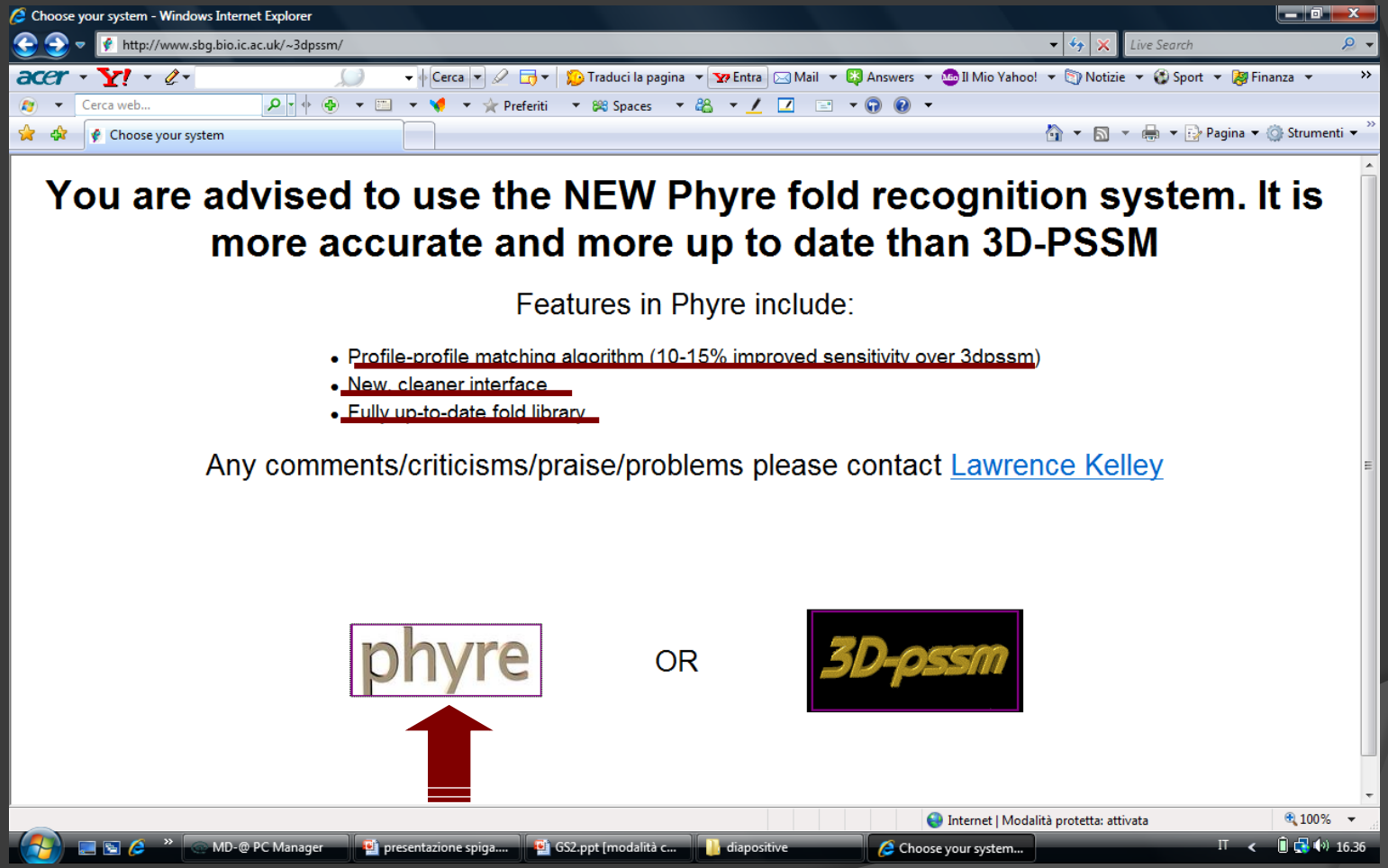
Per cui l'amminoacido nella sequenza sarà sostituito da una lettera corrispondente ad una delle 18 combinazioni

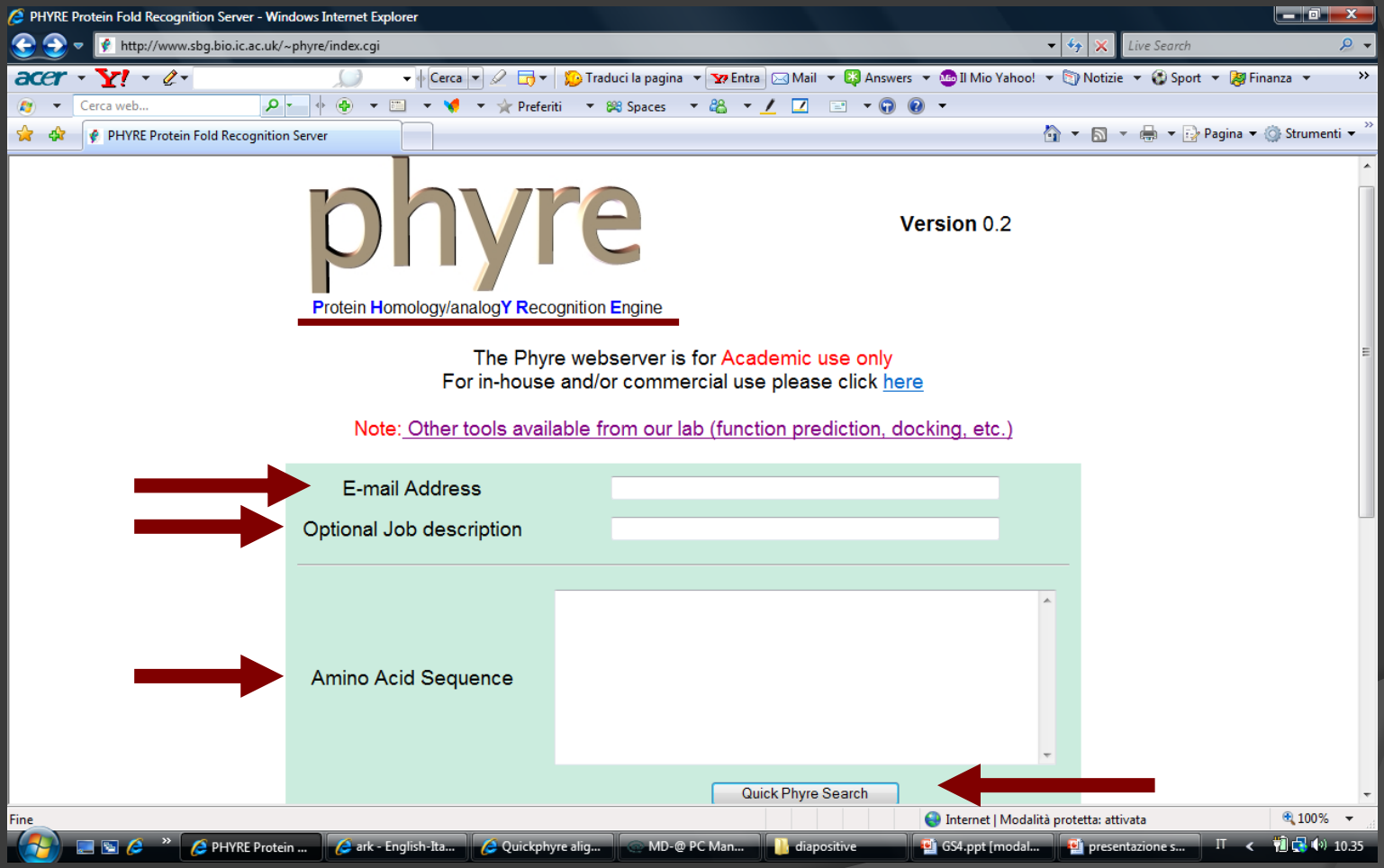
Se l'operazione viene applicata per tutte le sequenze di struttura nota la banca dati diventerà lineare

Metodi di THREADING=infilare

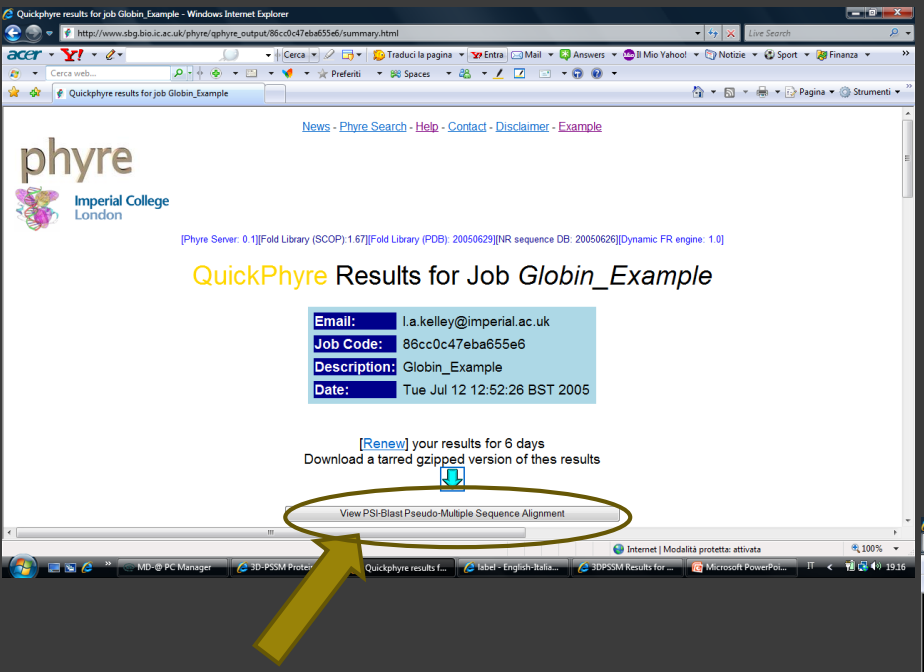
Questi metodi fanno qualcosa di molto simile a infilare una sequenza nella catena principale delle strutture note. Si ottengono così tanti possibili modelli della proteina usando come templato le proteine di struttura nota. I passi successivi sono la valutazione della “bontà” dei modelli approssimati e la selezione di uno o più con punteggio minore da un punto di vista energetico.


ExPASy - Tools - Windows Internet Explorer
http://www.expasy.org/tools/
ExPASy - Tools
Seq2Struct - A web resource for the identification of sequence-structure links
STRAP - A structural alignment program for proteins
TLSMD - TLS (Translation/Libration/Screw) Motion Determination
Tertiary structure prediction
Comparative modeling
SWISS-MODEL - An automated knowledge-based protein modelling server
3Djigsaw - Three-dimensional models for proteins based on homologues of known structure
CPHmodels - Automated neural-network based protein modelling server
ESyPred3D - Automated homology modeling program using neural networks
Geno3d - Automatic modelling of protein three-dimensional structure
SDSC1 - Protein Structure Homology Modeling Server
Threading
3D-PSSM - Protein fold recognition using 1D and 3D sequence profiles coupled with secondary structure information (Foldfit)
Fugue - Sequence-structure homology recognition
HHpred - Protein homology detection and structure prediction by HMM-HMM comparison
Libellula - Neural network approach to evaluate fold recognition results
LOOPP - Sequence to sequence, sequence to structure, and structure to structure alignment
SAM-T02 - HMM-based Protein Structure Prediction
Threader - Protein fold recognition
ProSup - Protein structure superimposition
SWEET - Constructing 3D models of saccharides from their sequences
Ab initio
HMMSTR/Rosetta - Prediction of protein structure from sequence
Assessing tertiary structure prediction
Anolea - Atomic Non-Local Environment Assessment
Internet | Modalità protetta: attivata
100%
MD-@ PC Manager
presentazione spiga....
GS2.ppt [modalità c...
diapositive
ExPASy - Tools - Wi...
16.39

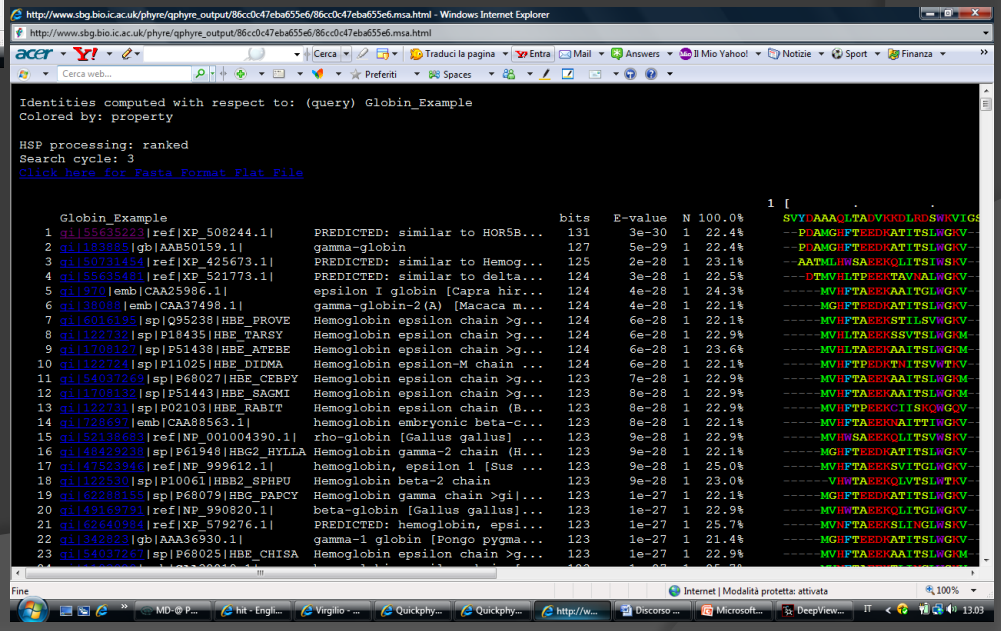

Choose your system - Windows Internet Explorer
http://www.sbg.bio.ic.ac.uk/~3dpssm/
Choose your system
You are advised to use the NEW Phyre fold recognition system. It is more accurate and more up to date than 3D-PSSM
Features in Phyre include:
Profile-profile matching algorithm (10-15% improved sensitivity over 3dpssm)
New, cleaner interface
Fully up-to-date fold library
Any comments/criticisms/praise/problems please contact Lawrence Kelley
phyre
OR
3D-pssm


PHYRE Protein Fold Recognition Server - Windows Internet Explorer
http://www.sbg.bio.ic.ac.uk/~phyre/index.cgi
PHYRE Protein Fold Recognition Server
phyre
Protein Homology/analogY Recognition Engine
Version 0.2
The Phyre webserver is for Academic use only
For in-house and/or commercial use please click here
Note: Other tools available from our lab (function prediction, docking, etc.)
E-mail Address
Optional Job description
Amino Acid Sequence
Quick Phyre Search
Fine
Internet | Modalità protetta: attivata
100%

ESEMPIO TIPICO DI RISPOSTA DI PHYRE

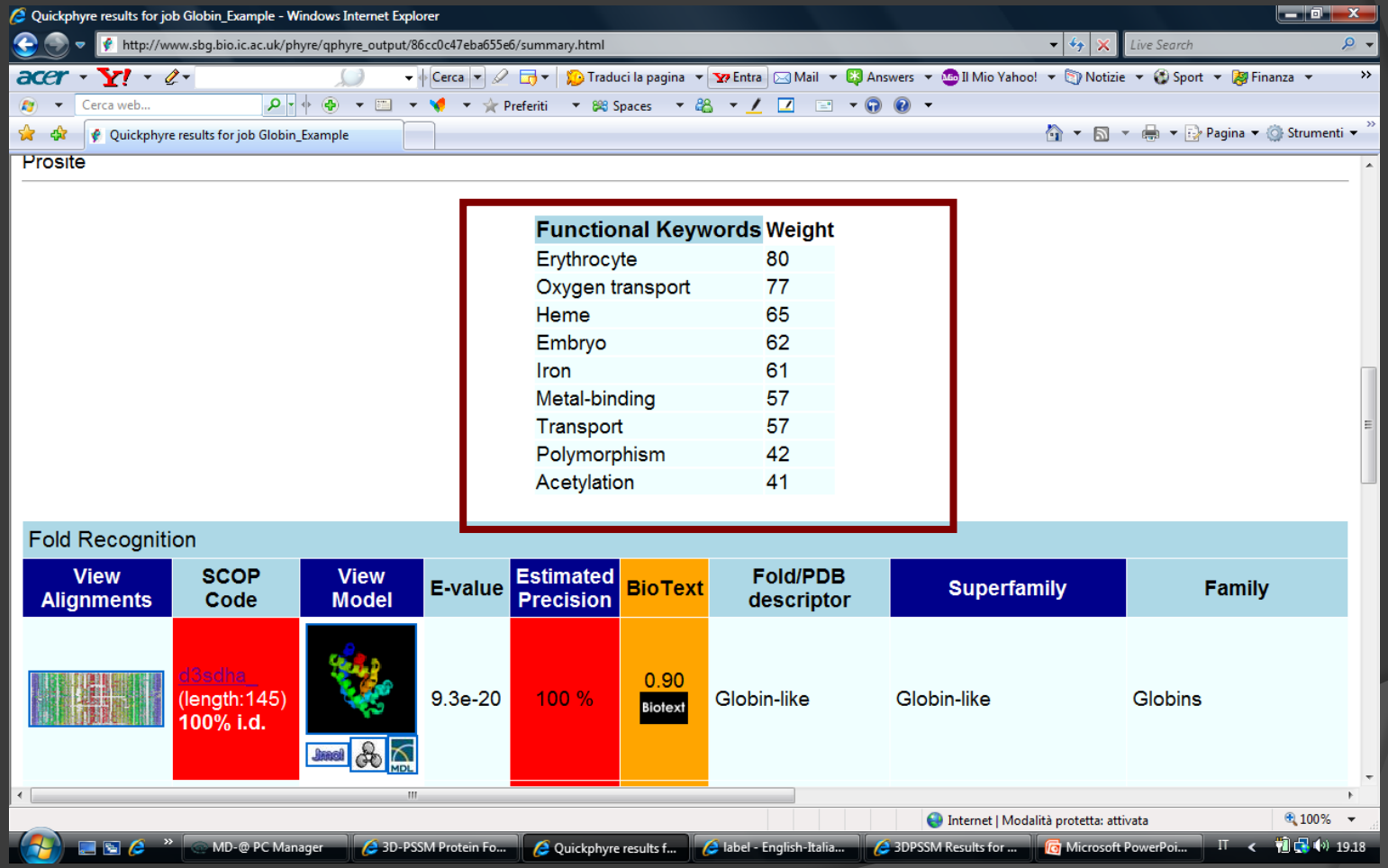


ALLINEAMENTI POSSIBILI FORNITI DA PSI-BLAST

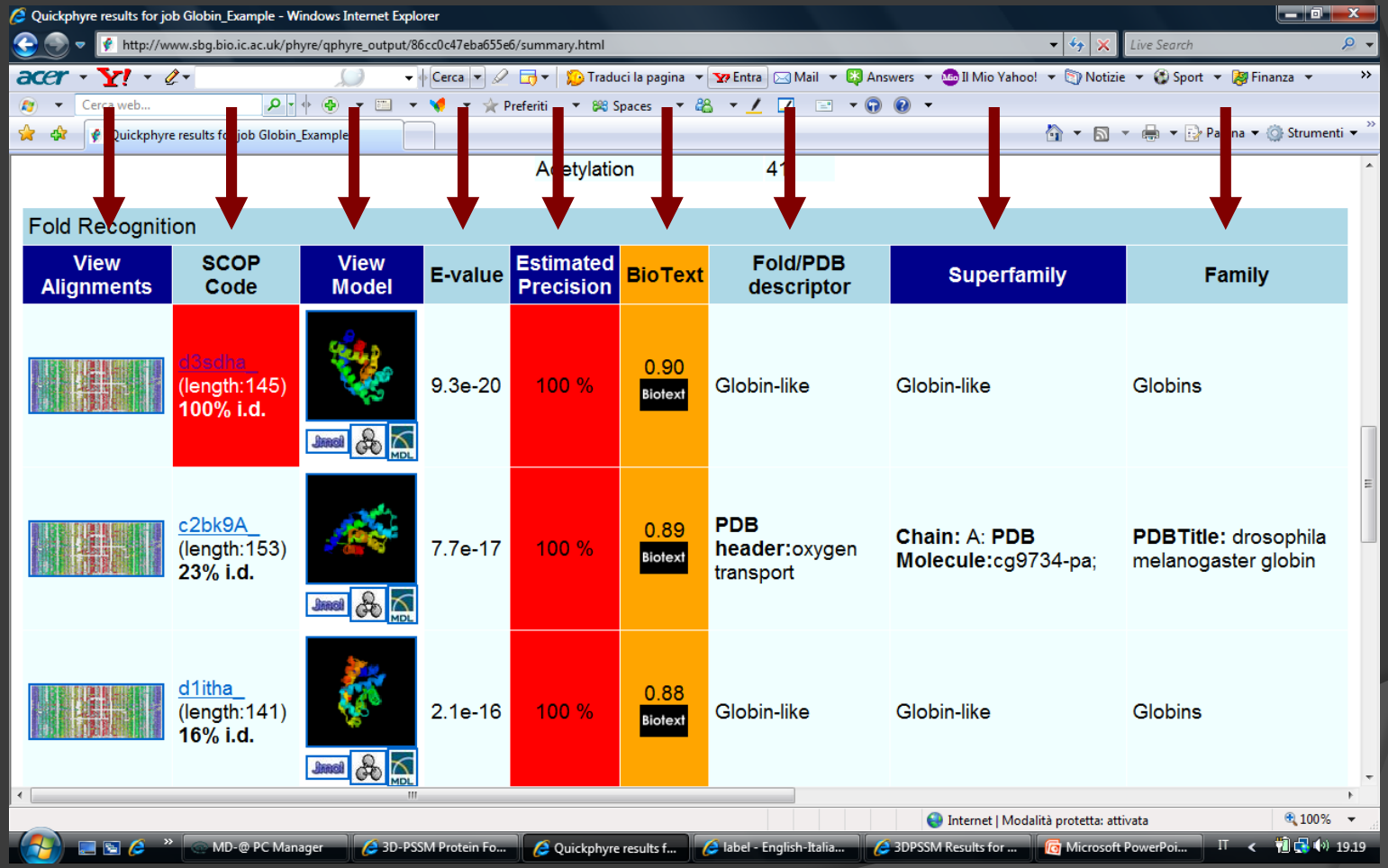


PRIMA ANALISI DELLA QUERY

PROBABILI FUNZIONI DELLA QUERY



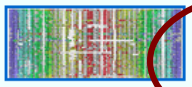
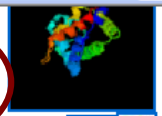
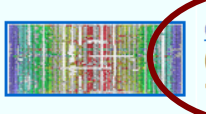
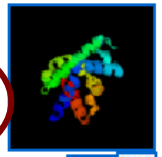
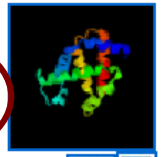
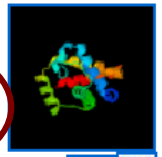
GLI ALLINEAMENTI CON MAGGIORE PUNTEGGIO



Quickphyre results for job Globin_Example - Windows Internet Explorer

http://www.sbg.bio.ic.ac.uk/phyre/qphyre_output/86cc0c47eba655e6/summary.html

Quickphyre results for job Globin_Example

	d1sctb_ (length:150) **55% i.d.**		1.8e-15	100 %	0.89 Biotext	Globin-like	Globin-like	Globins
	d1g08a_ (length:141) **19% i.d.**		2.5e-15	100 %	0.94 Biotext	Globin-like	Globin-like	Globins
	d1qpwa_ (length:141) **18% i.d.**		2.8e-15	100 %	0.94 Biotext	Globin-like	Globin-like	Globins
	d2gdm_ (length:153) **13% i.d.**		2.8e-15	100 %	0.84 Biotext	Globin-like	Globin-like	Globins

Internet | Modalità protetta: attivata

100%

MD-@ PC Manager | 3D-PSSM Protein Fo... | Quickphyre results f... | label - English-Italia... | 3DPSSM Results for ... | Microsoft PowerPoi...

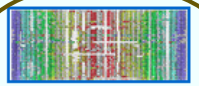
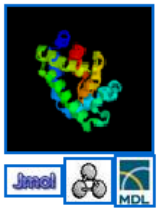
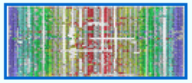
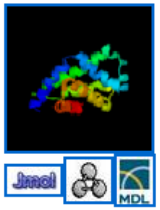
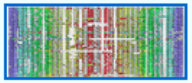
IT 19.20

Quickphyre results for job Globin_Example - Windows Internet Explorer

http://www.sbg.bio.ic.ac.uk/phyre/qphyre_output/86cc0c47eba655e6/summary.html

Acetylation 41

Fold Recognition

View Alignments	SCOP Code	View Model	E-value	Estimated Precision	BioText	Fold/PDB descriptor	Superfamily	Family
	d3sdha_ (length:145) **100% i.d.**		9.3e-20	100 %	0.90 Biotext	Globin-like	Globin-like	Globins
	c2bk9A_ (length:153) **23% i.d.**		7.7e-17	100 %	0.89 Biotext	**PDB header:**oxygen transport	**Chain:** A: **PDB Molecule:**cg9734-pa;	**PDBTitle:** drosophila melanogaster globin
	d1itha_ (length:141) **16% i.d.**		2.1e-16	100 %	0.88 Biotext	Globin-like	Globin-like	Globins

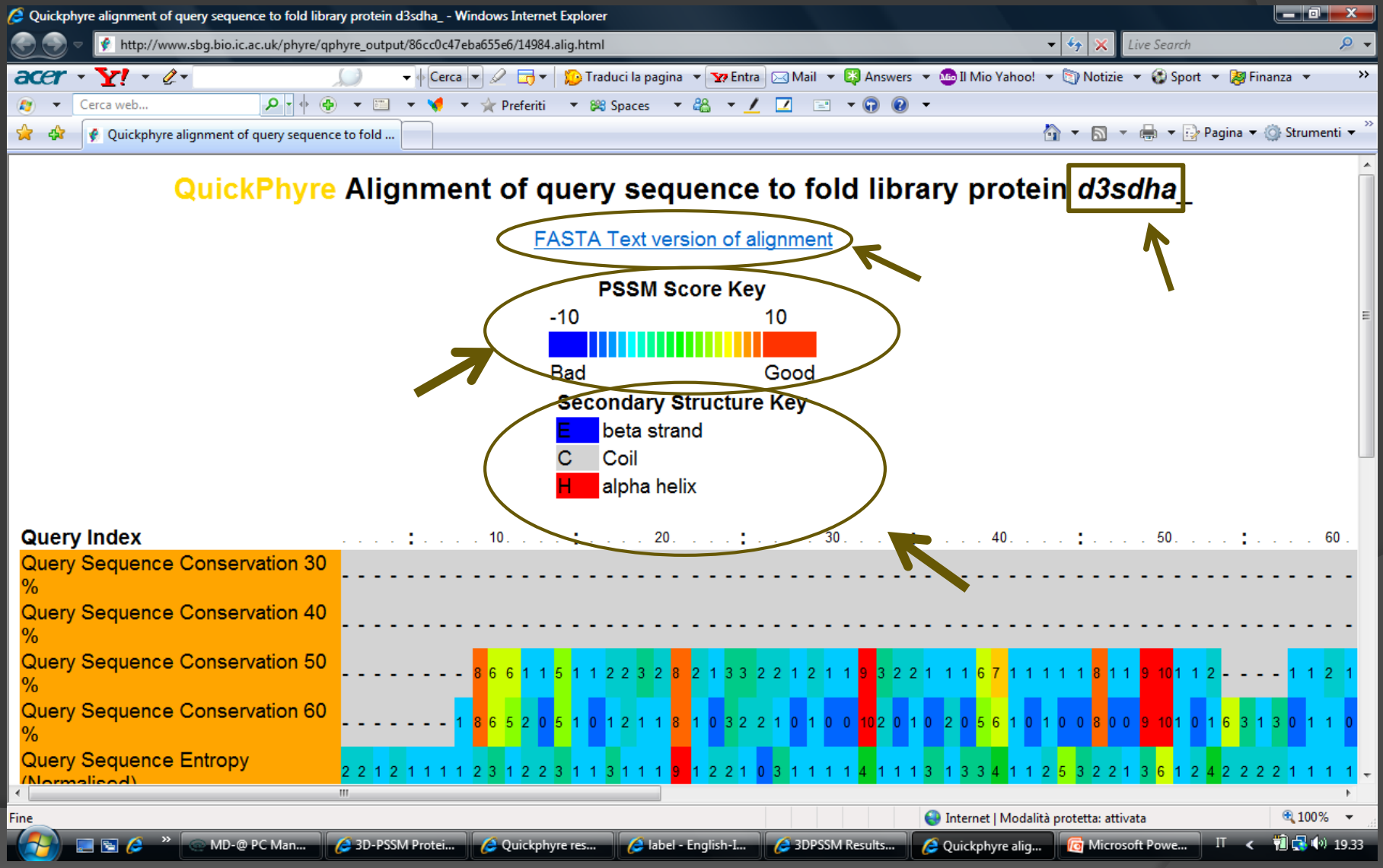
Internet | Modalità protetta: attivata

100%

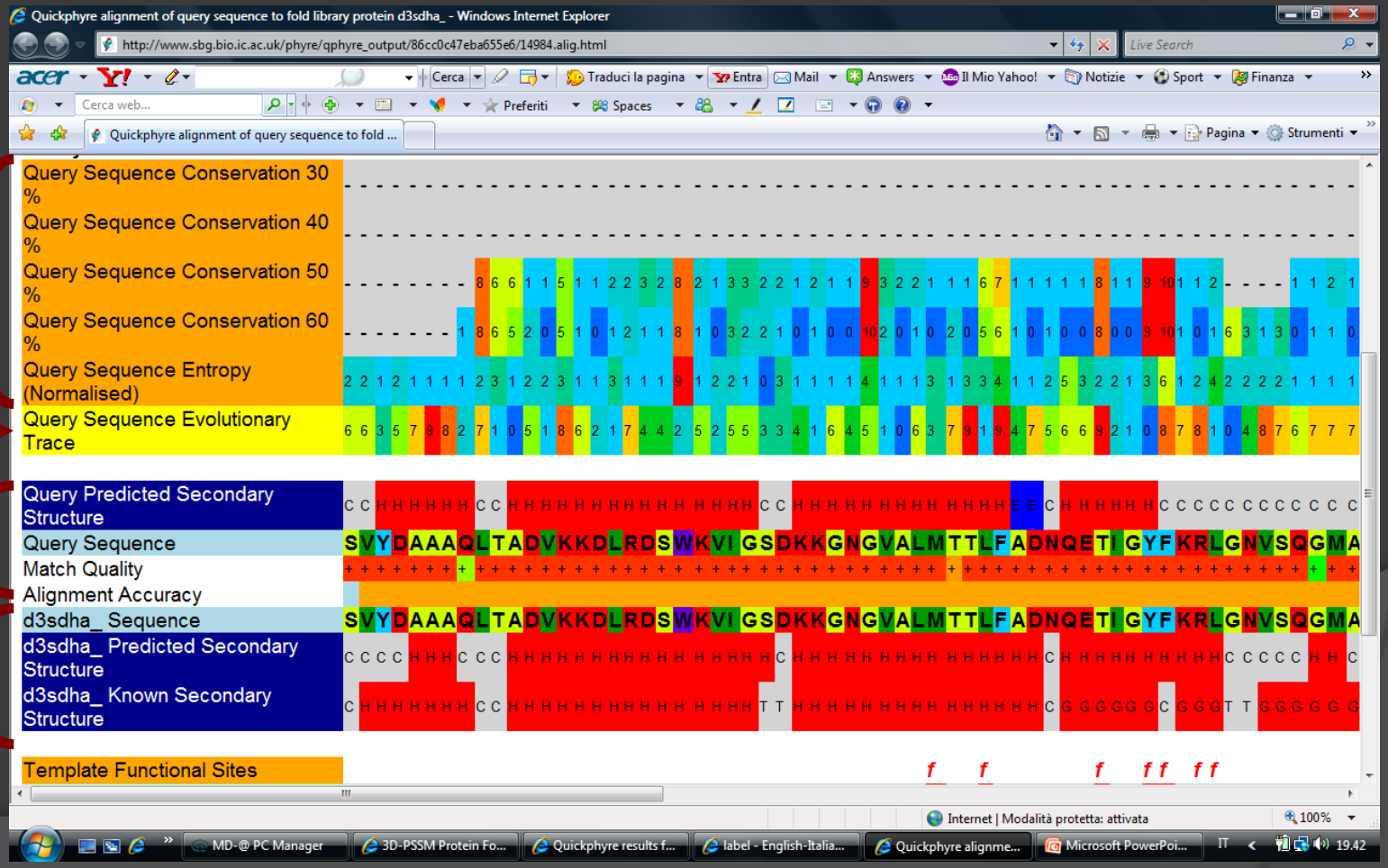
MD-@ PC Manager | 3D-PSSM Protein Fo... | Quickphyre results f... | label - English-Italia... | 3DPSSM Results for ... | Microsoft PowerPoi...

IT 19.19

ALLINEAMENTO MIGLIORE

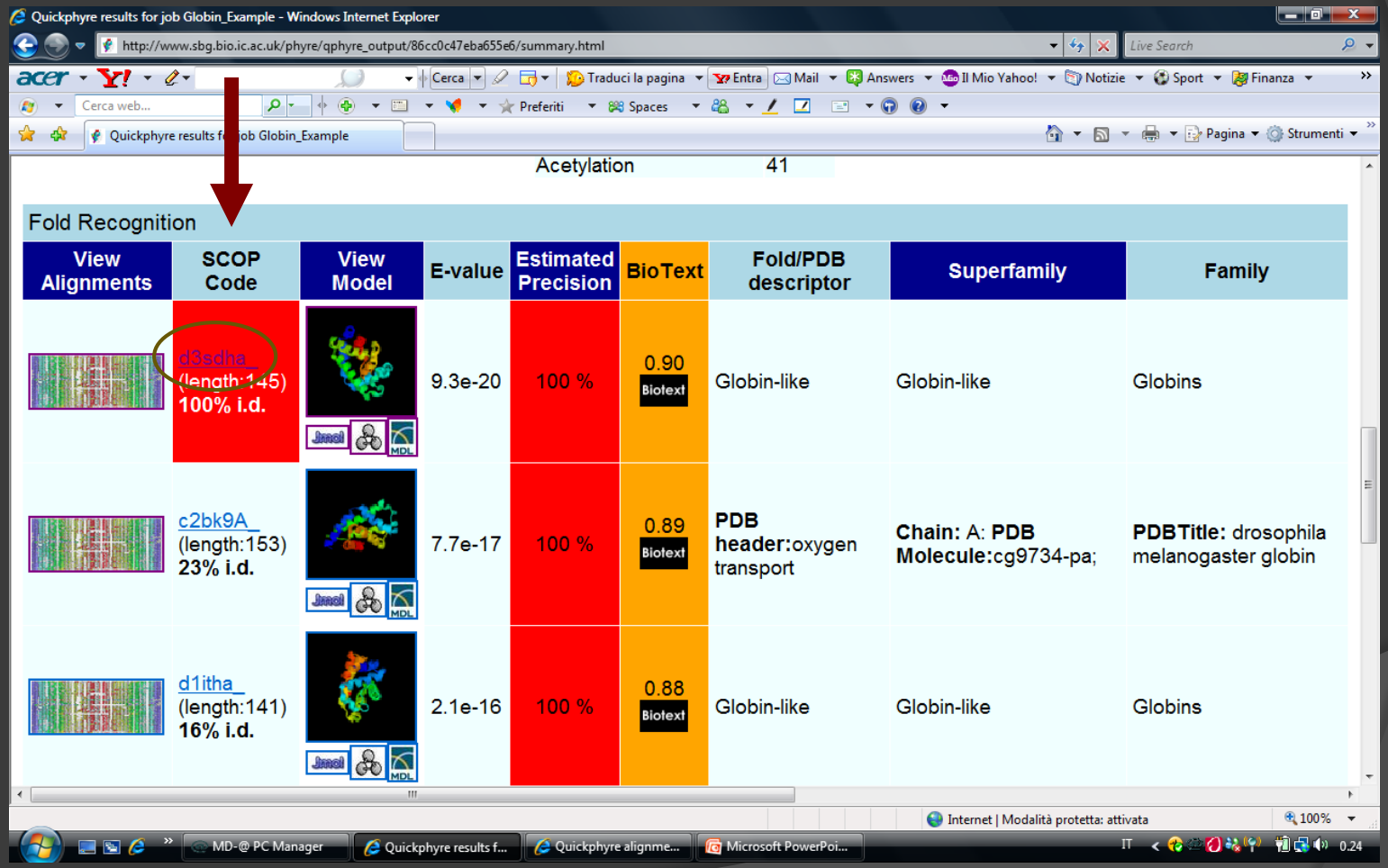


CONSERVAZIONE, TRACCIA EVOLUTIVA E PREDIZIONE DI STRUTTURA SECONDARIA

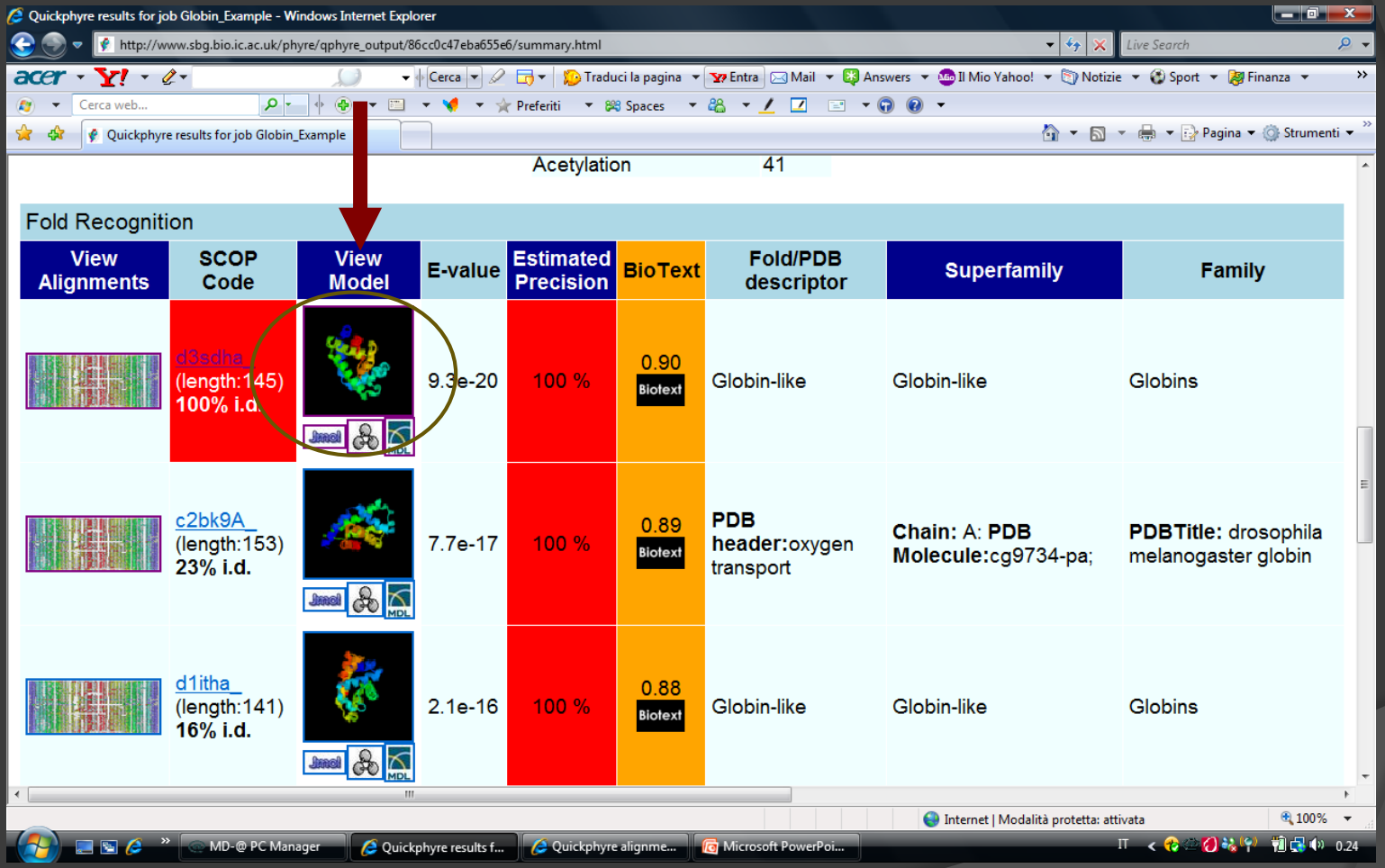


STRUTTURA SECONDARIA E SITI FUNZIONALI

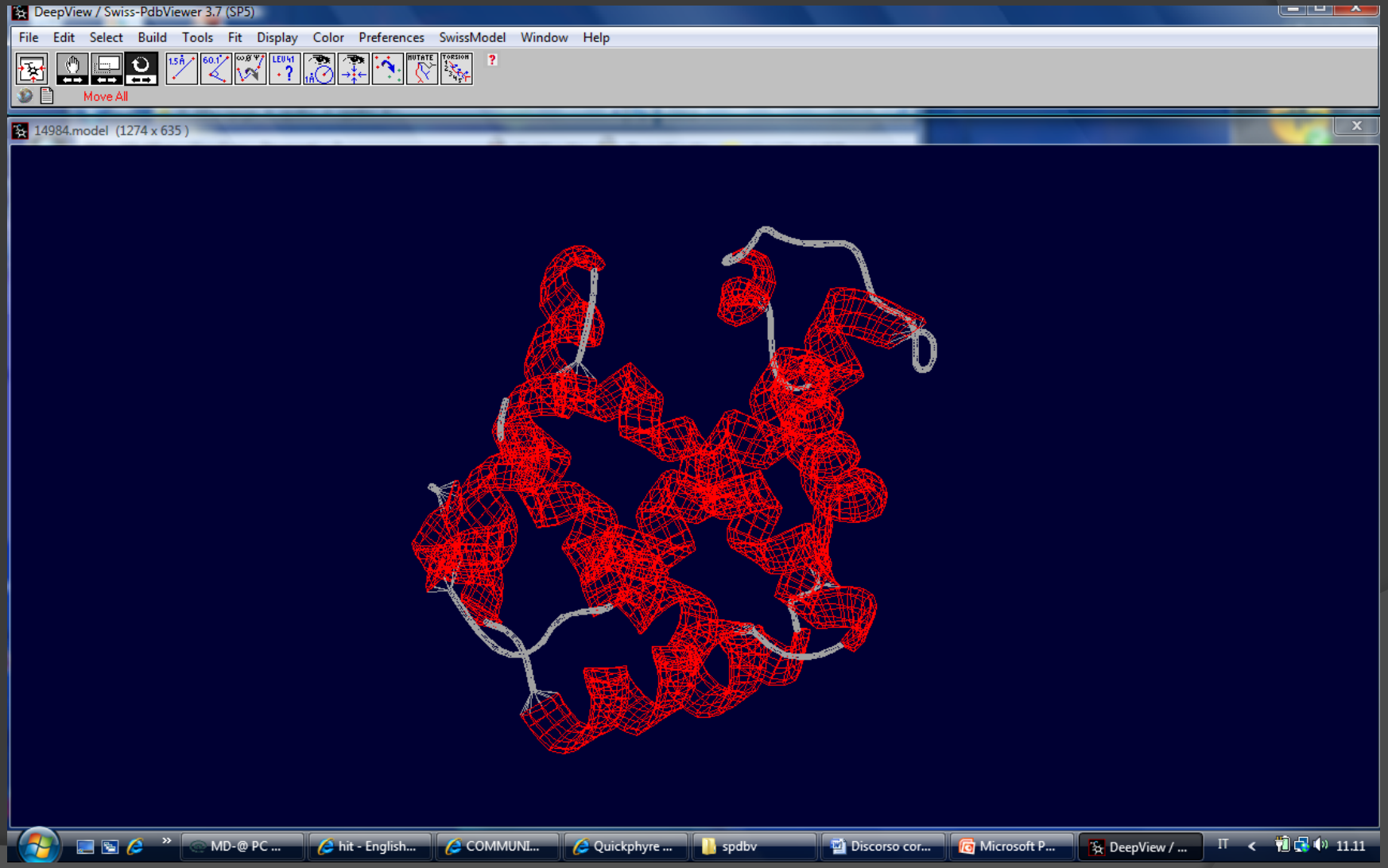
CODICE SCOP = Structural Classification Of Proteins



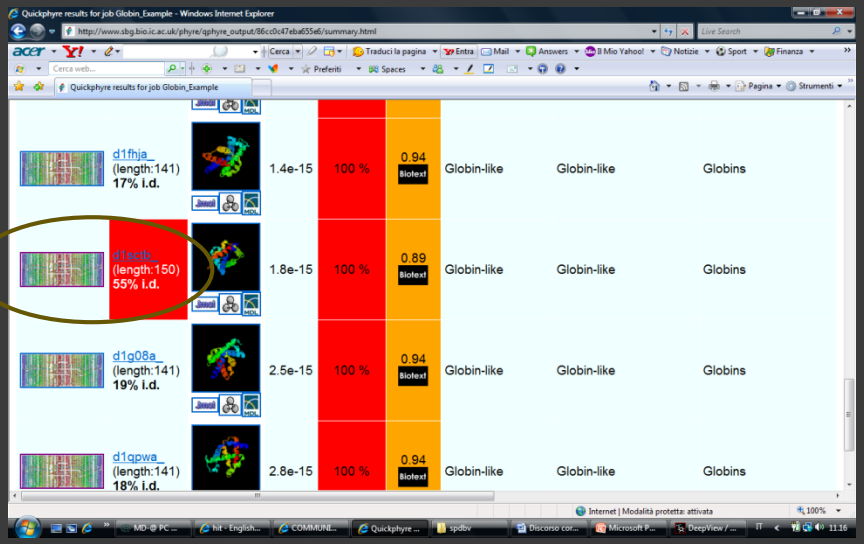
View Model


Quickphyre results for job Globin_Example - Windows Internet Explorer
http://www.sbg.bio.ic.ac.uk/phyre/qphyre_output/86cc0c47eba655e6/summary.html
Quickphyre results for job Globin_Example
Acetylation 41
Fold Recognition
View Alignments
SCOP Code
View Model
E-value
Estimated Precision
BioText
Fold/PDB descriptor
Superfamily
Family
d3sdha_ (length:145) 100% i.d.
9.3e-20
100 %
0.90 Biotext
Globin-like
Globin-like
Globins
c2bk9A_ (length:153) 23% i.d.
7.7e-17
100 %
0.89 Biotext
PDB header:oxygen transport
Chain: A: PDB Molecule:cg9734-pa;
PDBTitle: drosophila melanogaster globin
d1itha_ (length:141) 16% i.d.
2.1e-16
100 %
0.88 Biotext
Globin-like
Globin-like
Globins
Internet | Modalità protetta: attivata
100%

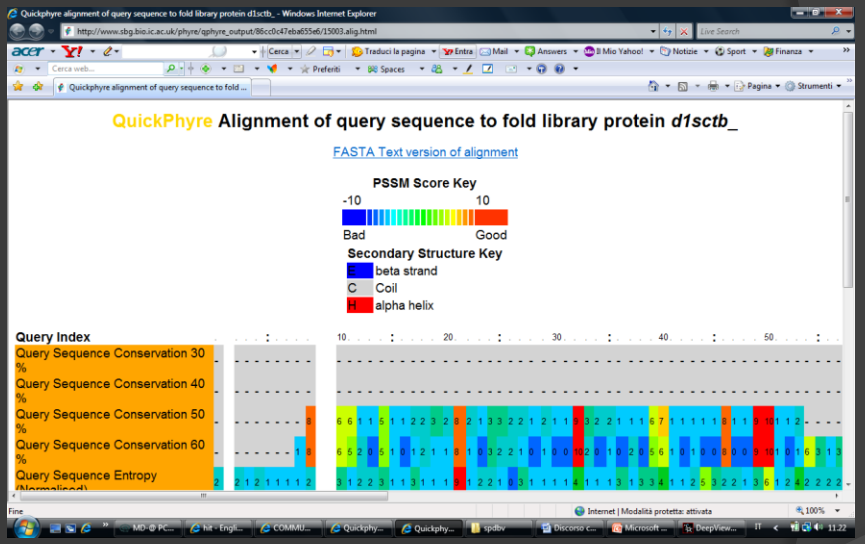
MODELLO DELLA d3sdh_a

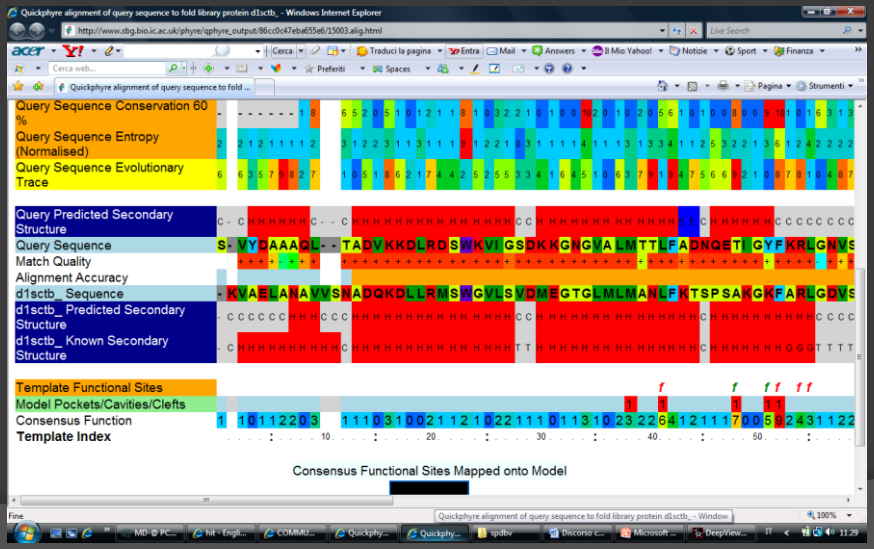


CONFRONTO CON PROTEINA COL 55% D'IDENTITA': d1sct_b

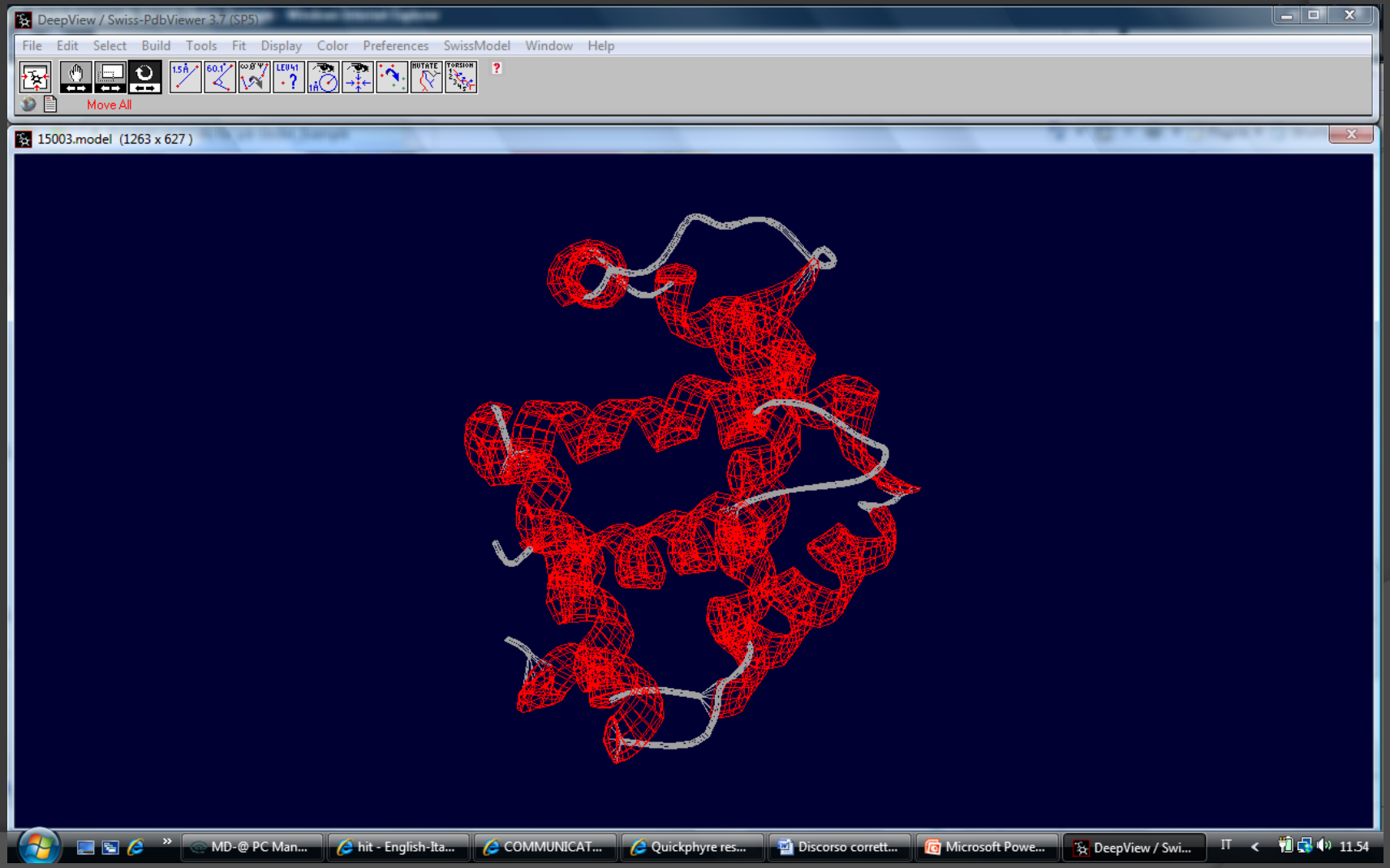


Vedi l'allineamento

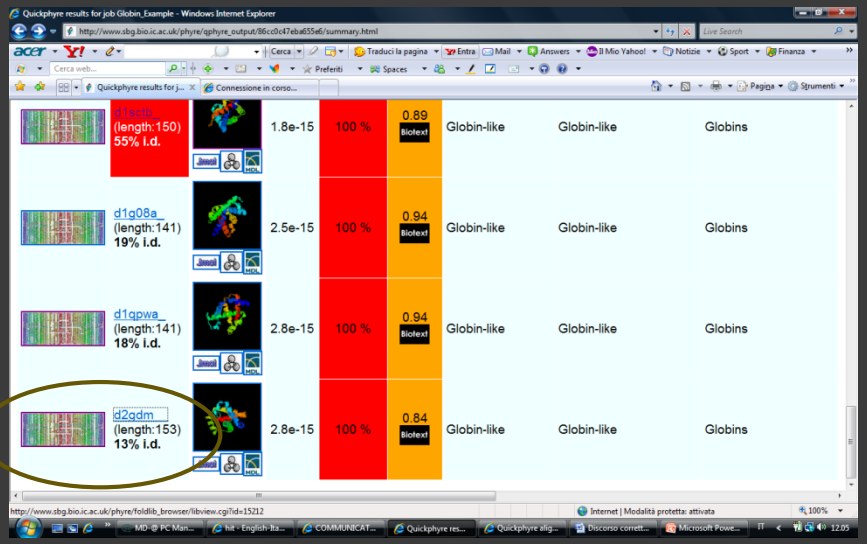




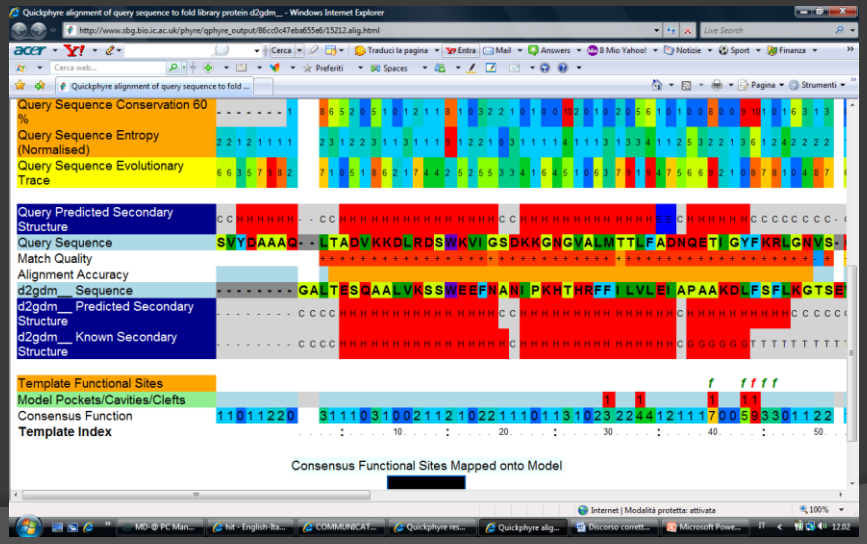
MODELLO OTTENUTO USANDO COME TEMPLATE d1sct_b



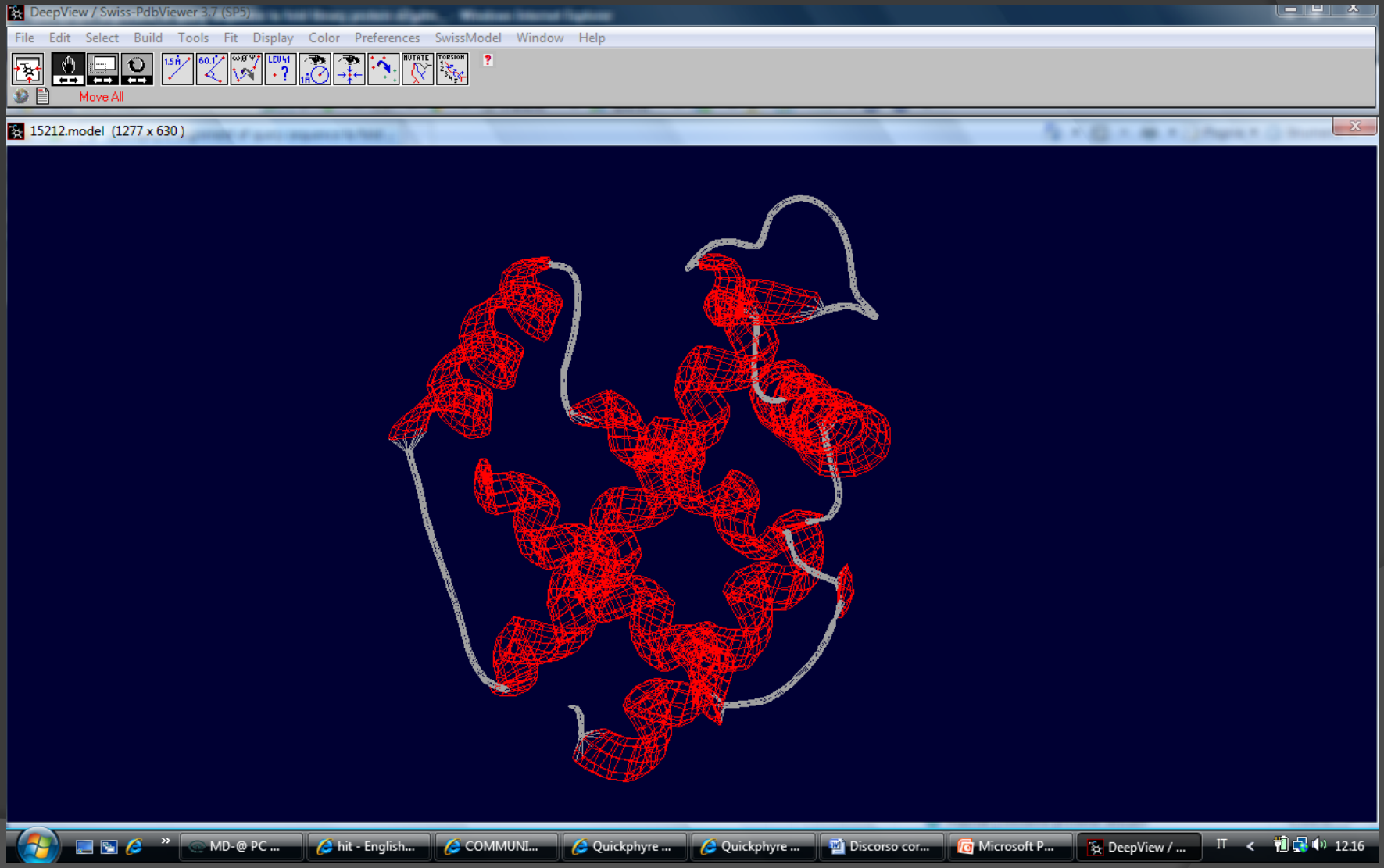
CONFRONTO CON PROTEINA COL 13% D'IDENTITA': d2gdm



Vedi l'allineamento



MODELLO OTTENUTO USANDO COME TEMPLATE d2gdm



Il benchmark di un metodo predittivo

Farlo da soli utilizzando i vari metodi richiederebbe molto tempo e spazio di calcolo.

EVA è un server che ogni giorno scarica dal PDB le nuove strutture ne estrapola la sequenza e prova usando i vari metodi predittivi ad ottenere una struttura usando similitudine di sequenza non superiore al 33% solo automatic servers

CASP (critical assessment techniques for protein structure prediction)

SARS 1

Comparsa nel 2003 di un nuovo virus	?
Sequenziamento del genoma completo	NC_004718
Identificazione della classe di appartenenza	Coronavirus ssRNA
Identificazione delle proteine espresse	14 proteine
Identificazione delle proteine responsabili dell'attacco alla cellula ospite	Spike protein
Identificazione della struttura delle proteine di interesse	S1 + S2

Cosa si può fare con la bioinformatica

Identificazione di una nuova specie

Sars CoVirus

Determinare in tempi brevi la struttura di alcune proteine

Spike protein
Peplomero
Elicasi

Drug designe
Diagnostic kits
Vaccine

Sequenza nucleotidica

Sequenza amminoacidica

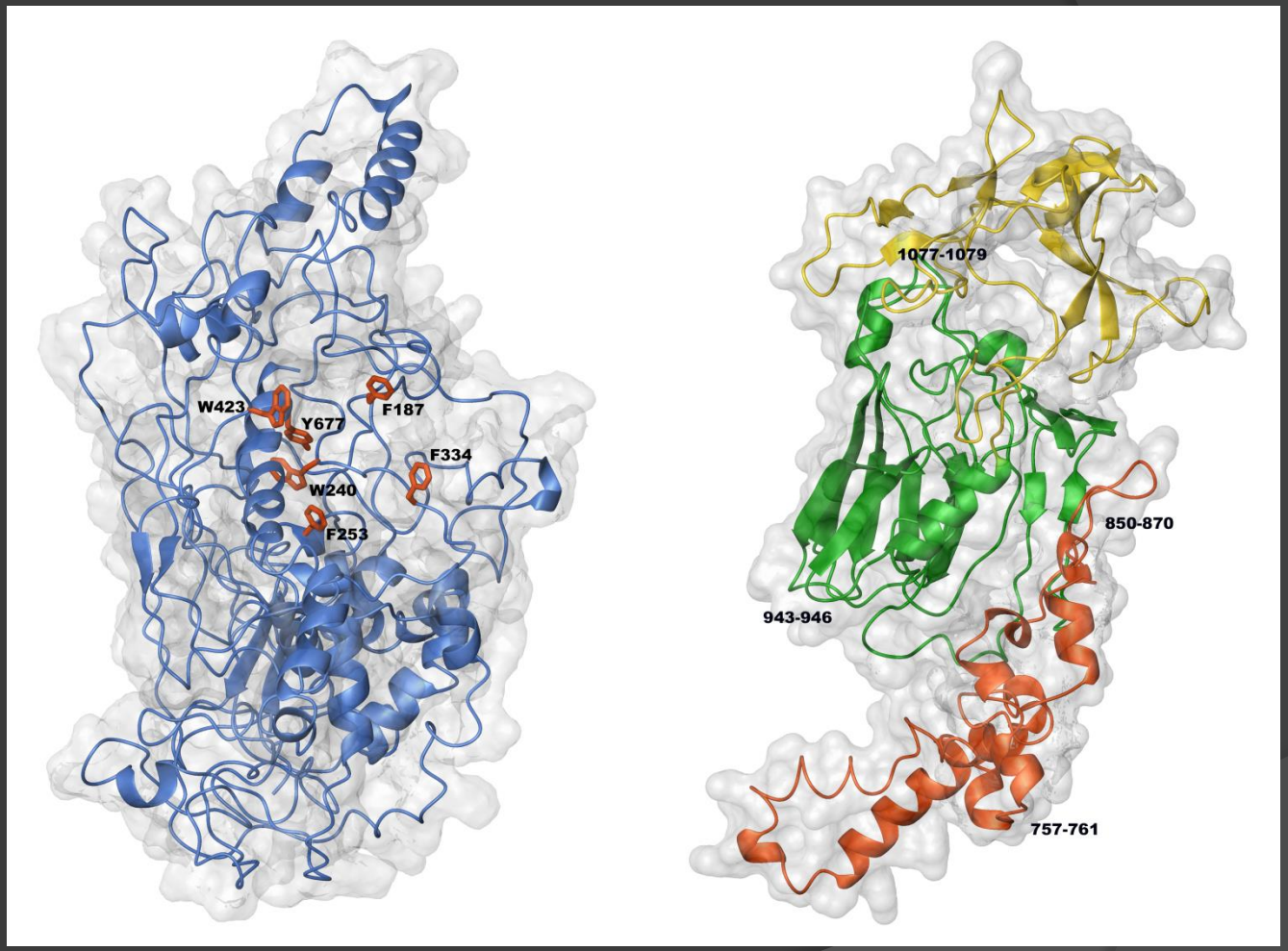
Metodi bioinformaticidi predizione strutturale

Modelling

Struttura Tridimensionale

1G9D neurotoxin B of *Clostridium botulinum*

15% identity
49% positivity
5% gaps



S1 1Q4Z

S2 1Q4Y

SARS 2

SEM Scanning Electron Microscopy

Struttura quaternaria

Trimero della spike protein sulla superficie

Ricostruzione del trimero con approccio bioinformatico

Struttura di S1 ed S2 precedentemente modellate

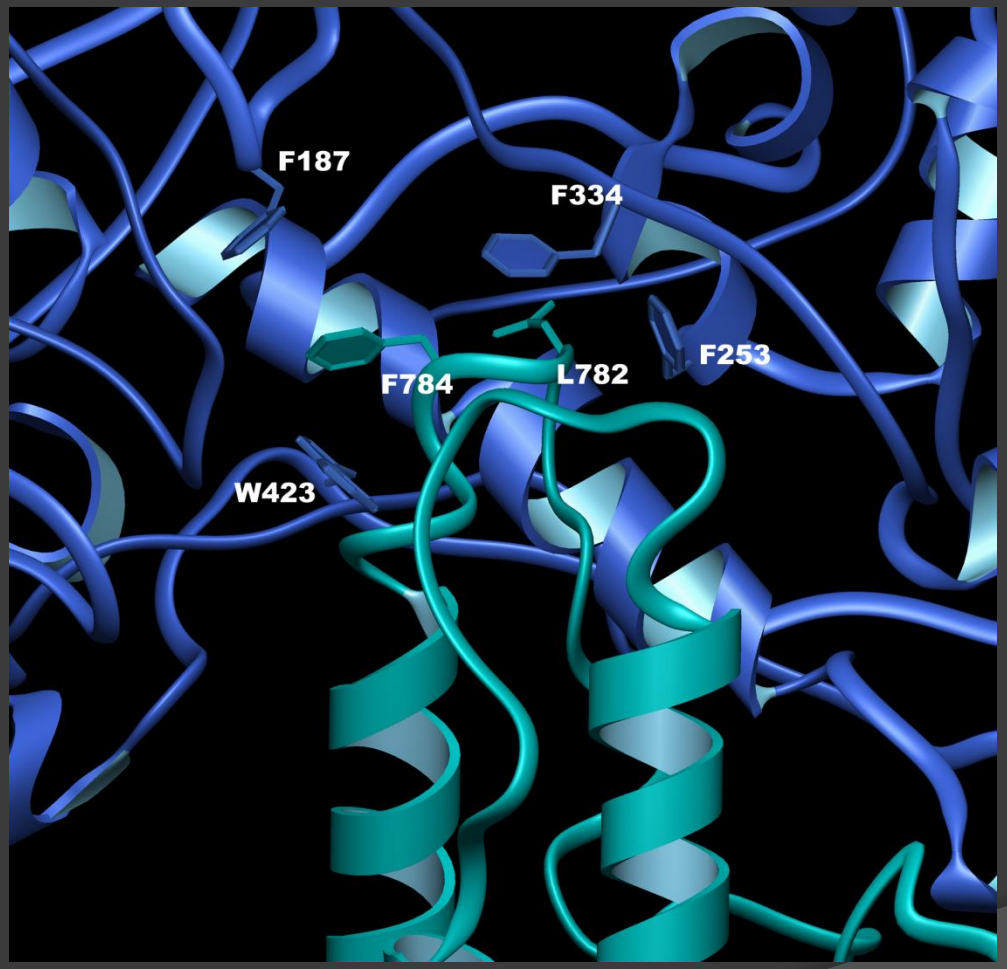
Docking molecolare tra le due subunità

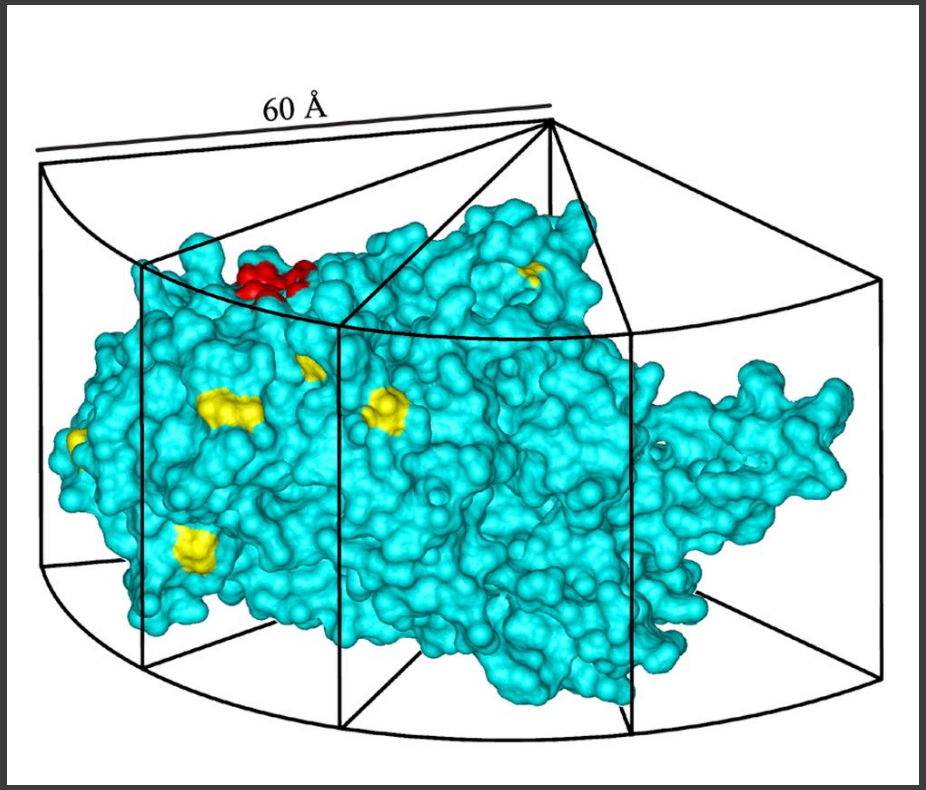
S1 testa globulare del peplomero con attività di legame al recettore cellulare
S2 gambo del peplomero

1 step: modelling del gambo di S2 coiled coil e HRregion

2 step: allineamento di 253 sequenze di S1
alcuni residui sono sempre conservati

3 step: tasca idrofobica necessaria per l'assemblaggio del peplomero


F187
F334
F253
L782
F784
W423

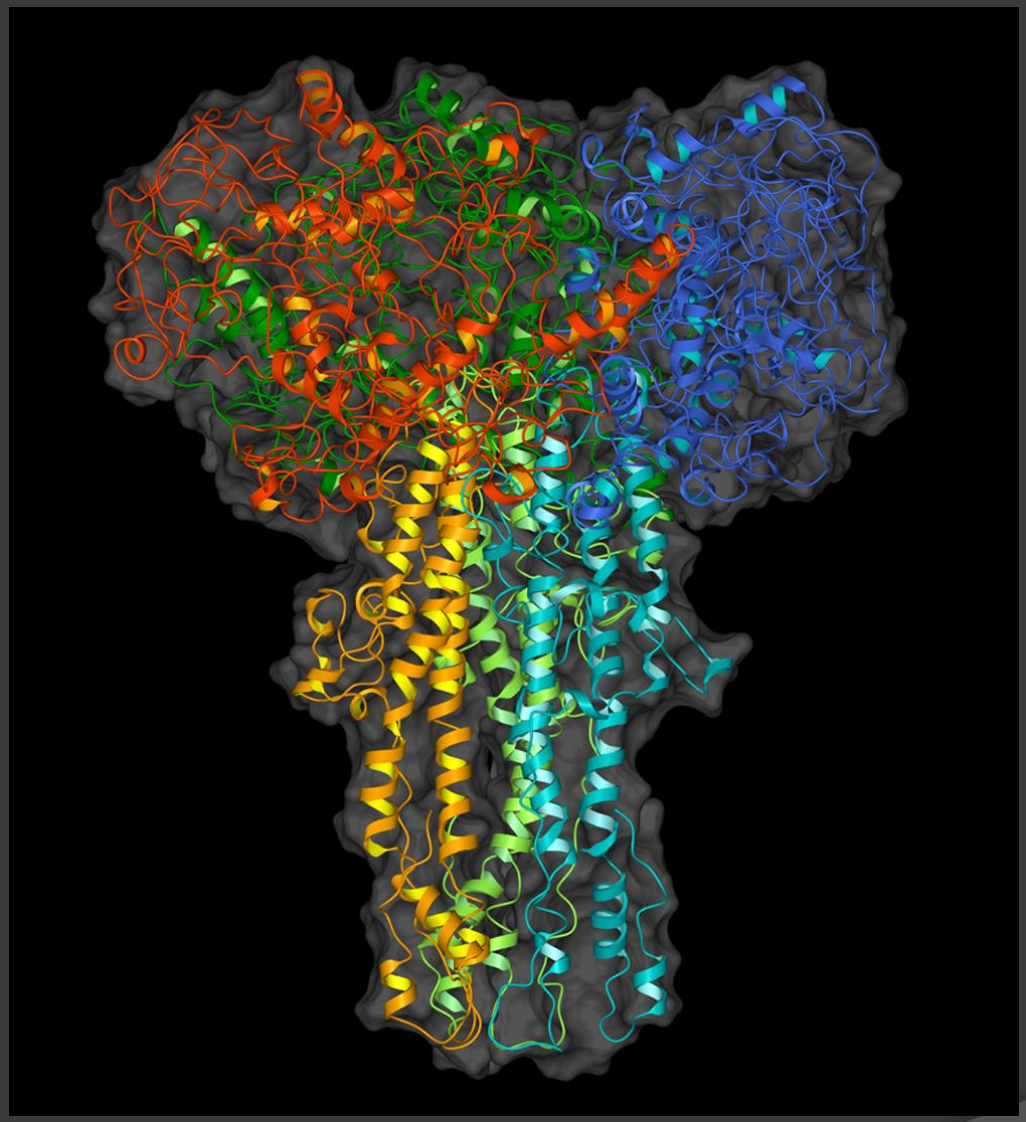


Ricostruzione del peplomero

Orientazione di S1+S2

Siti glicosilazione esterni

Siti idrofobici interni



Peplomer 1T7G Brevettato per potenziali farmaci

SARS 3

SCV helicase inibirne il funzionamento per impedire al virus di replicarsi

Drug designe: necessario conoscere la struttura terziaria

NP_828870 sequenza amminoacidica

Modelling

Template 1UAA Rep helicase of E. coli

L’analisi della sequenza mostrava due domini funzionali

Metal Binding Domain + Hel domain

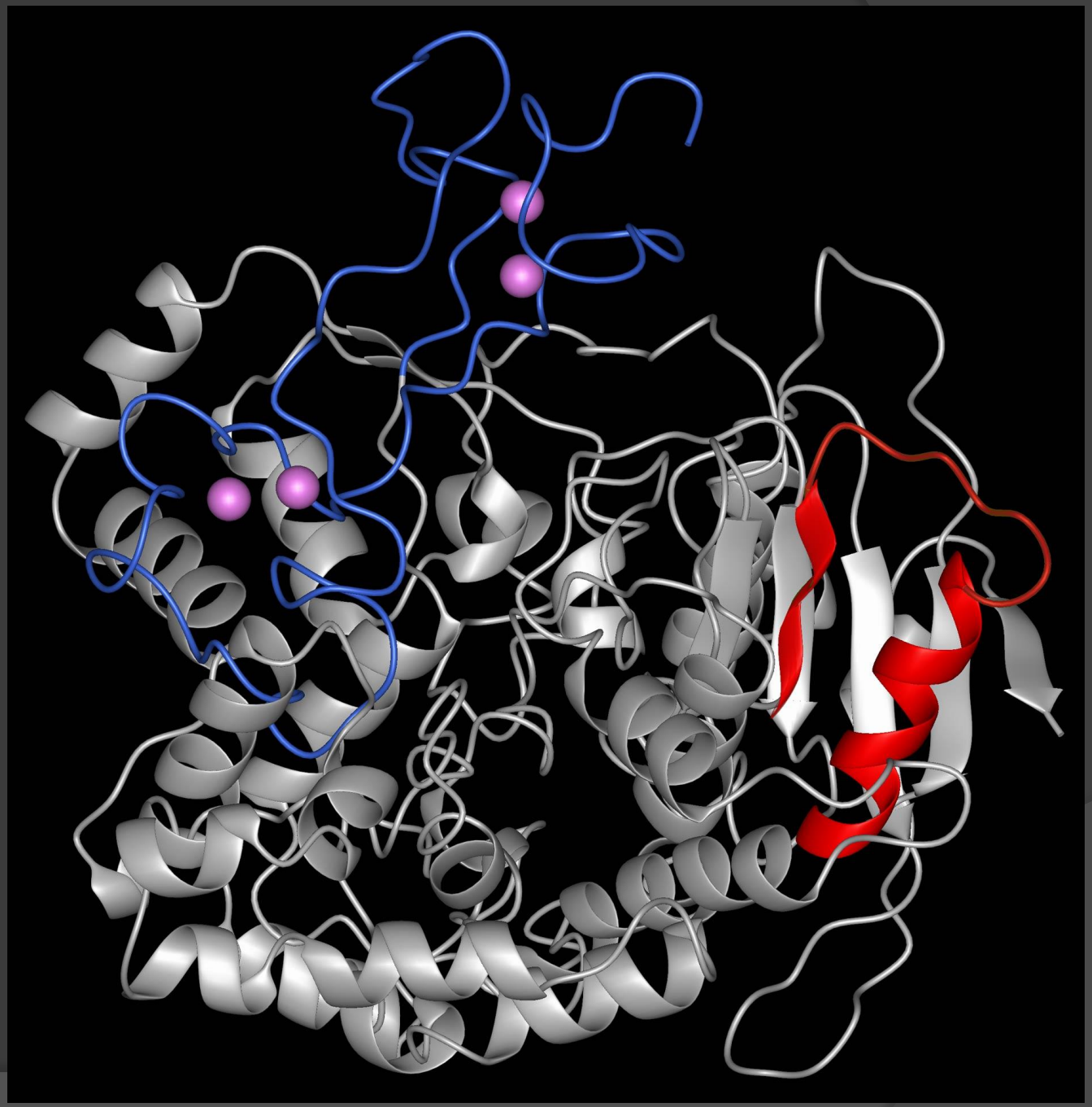
MBD ricco di cisteine e istidine complessante lo zinco > inibita dal bismuto > come ?

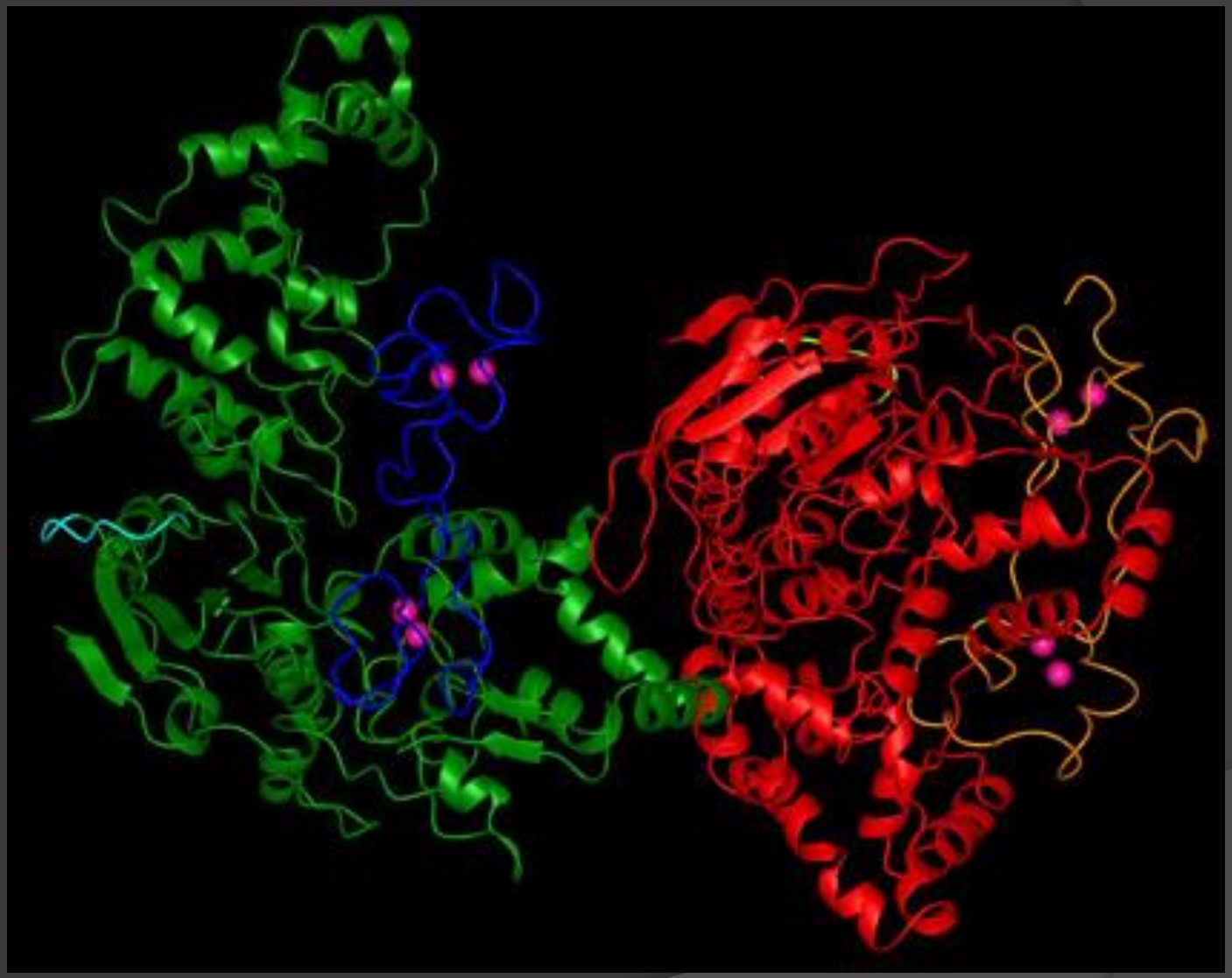
Confronto con altre elicasi che presentavano un MBD

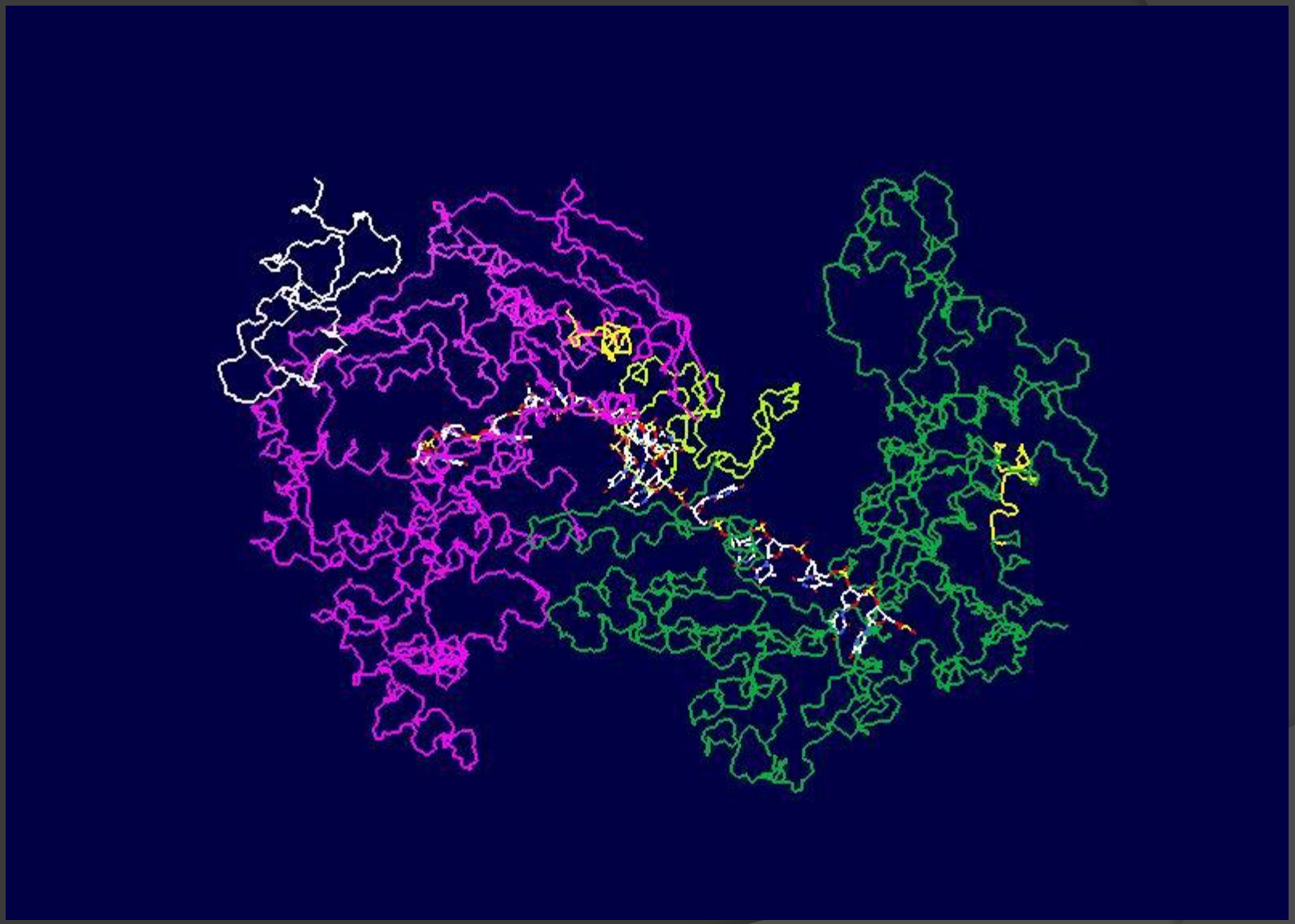
Arterivirus helicase possiedono un MBD complessante 4 atomi di zinco

Possibili modelli presenti in altre proteine
Zn2Cys6 GAL4-like protein
Zn2Cys4His2 RAG1 domain

Ricostruzione della struttura geometrica tetraedrica







ALBUMINA

Albumina umana > cristallo

Albumina bovina > modello

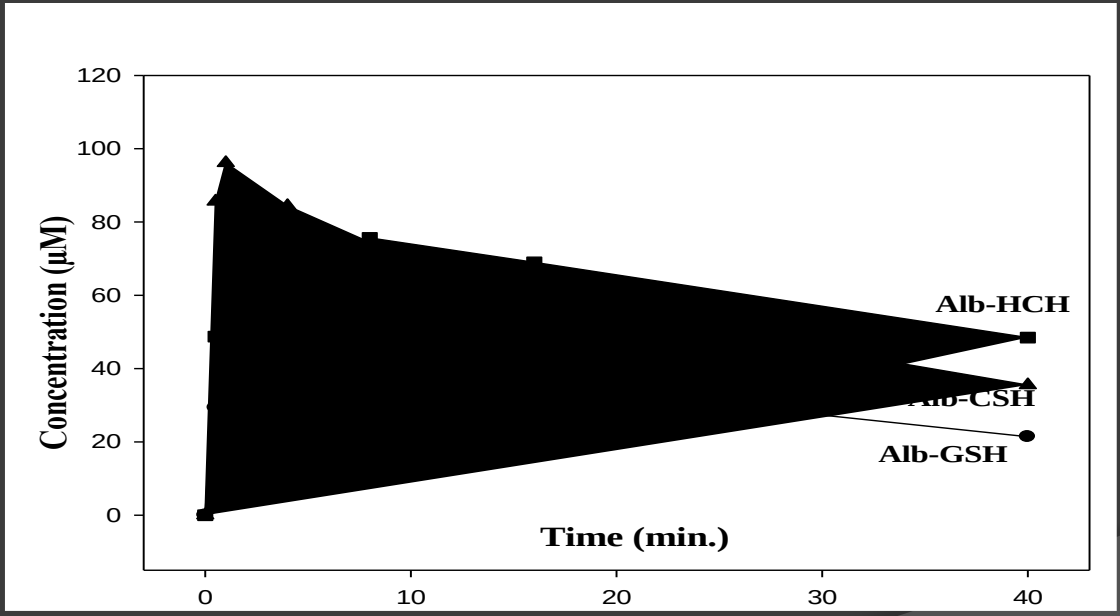
Albumina ratto > modello

Albumina suino > modello

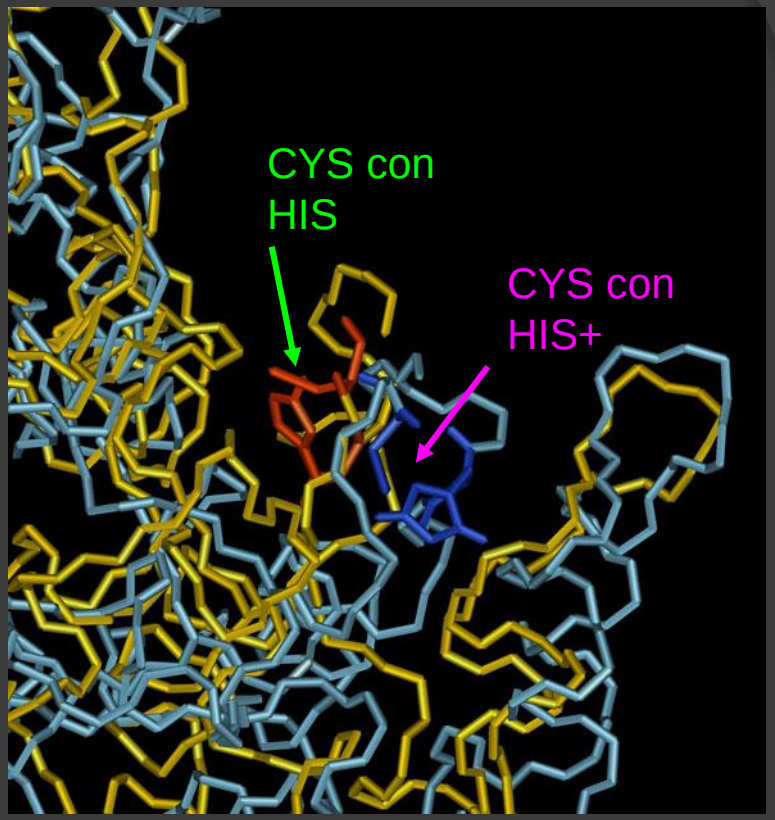
Valutazione sperimentale della loro reattività nei confronti di diversi tioli

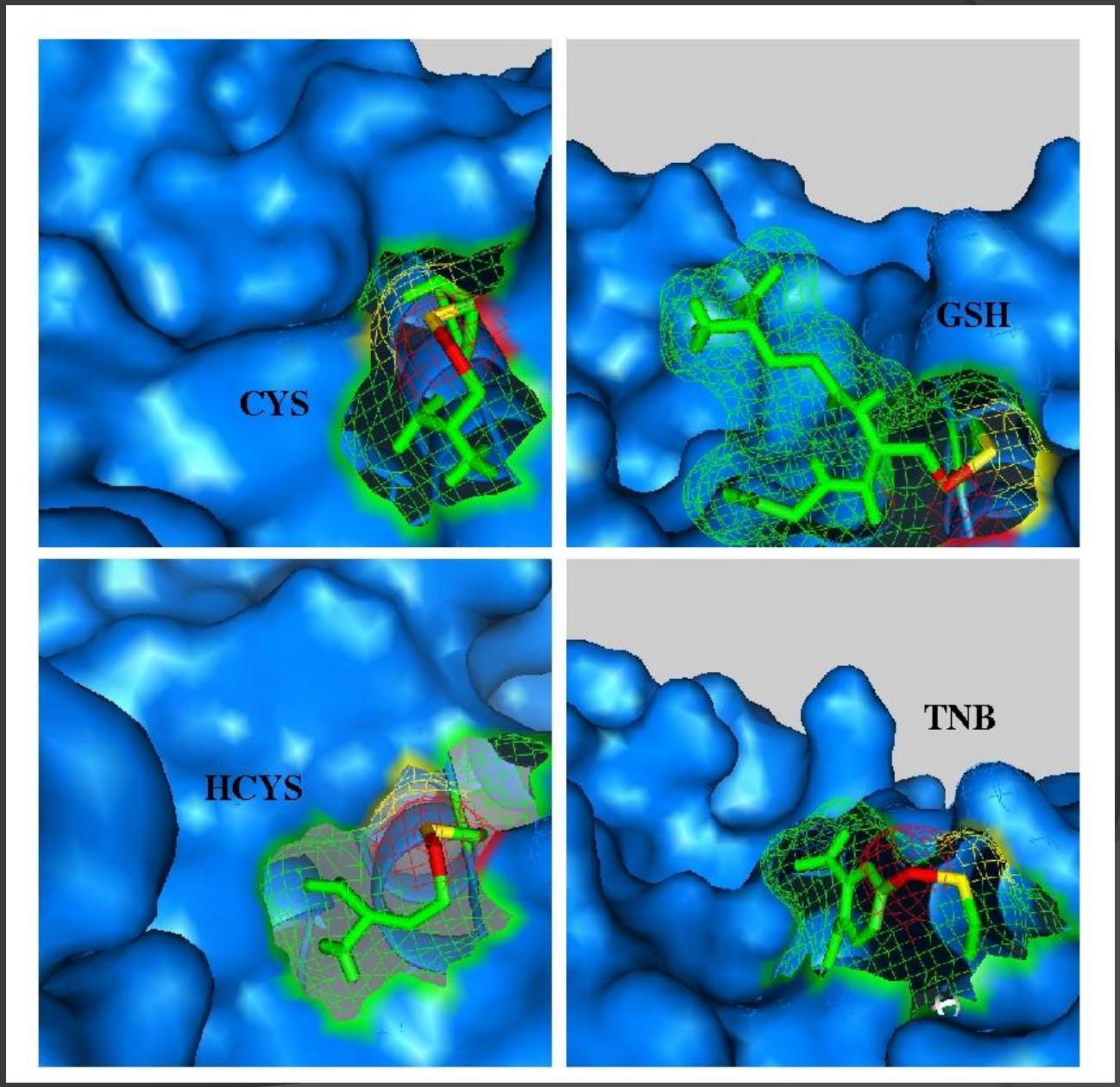
Sequenze molto simili spiegazioni attraverso le strutture

Valutazione sperimentale della loro reattività con diversi tioli:
Glutatione
Cistina
Omocisteina

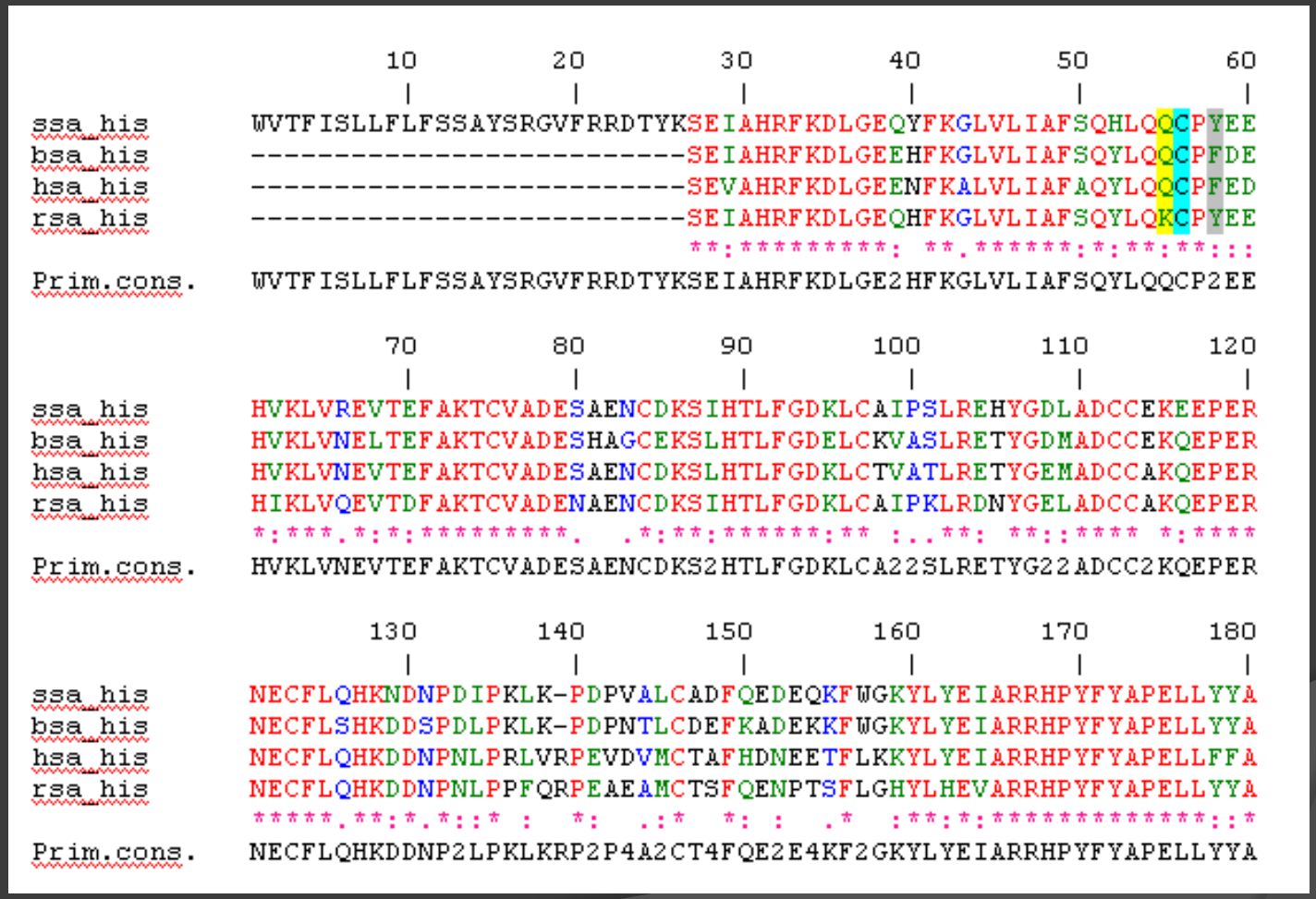


Variazioni del pH influenzano l’esposizione della cisteina Flip-Flop




CYS
GSH
HCYS
TNB

Quali sono le differenze strutturali che spieganio il differente comportamento delle 4 albumine?



ADP

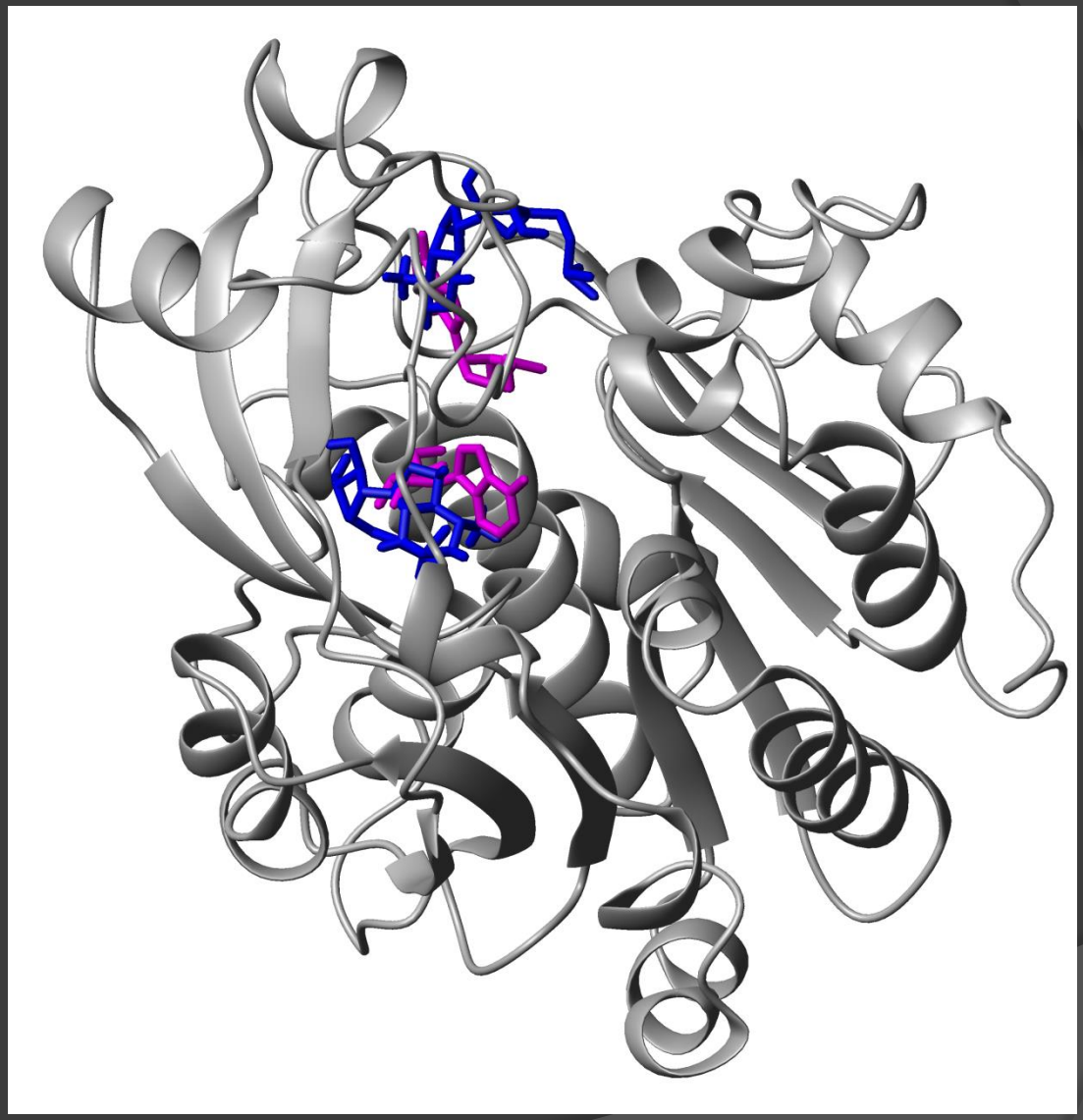
Due proteine:

Adenosina Kinasi che lega ATP+Adenosine dà ADP +AMP

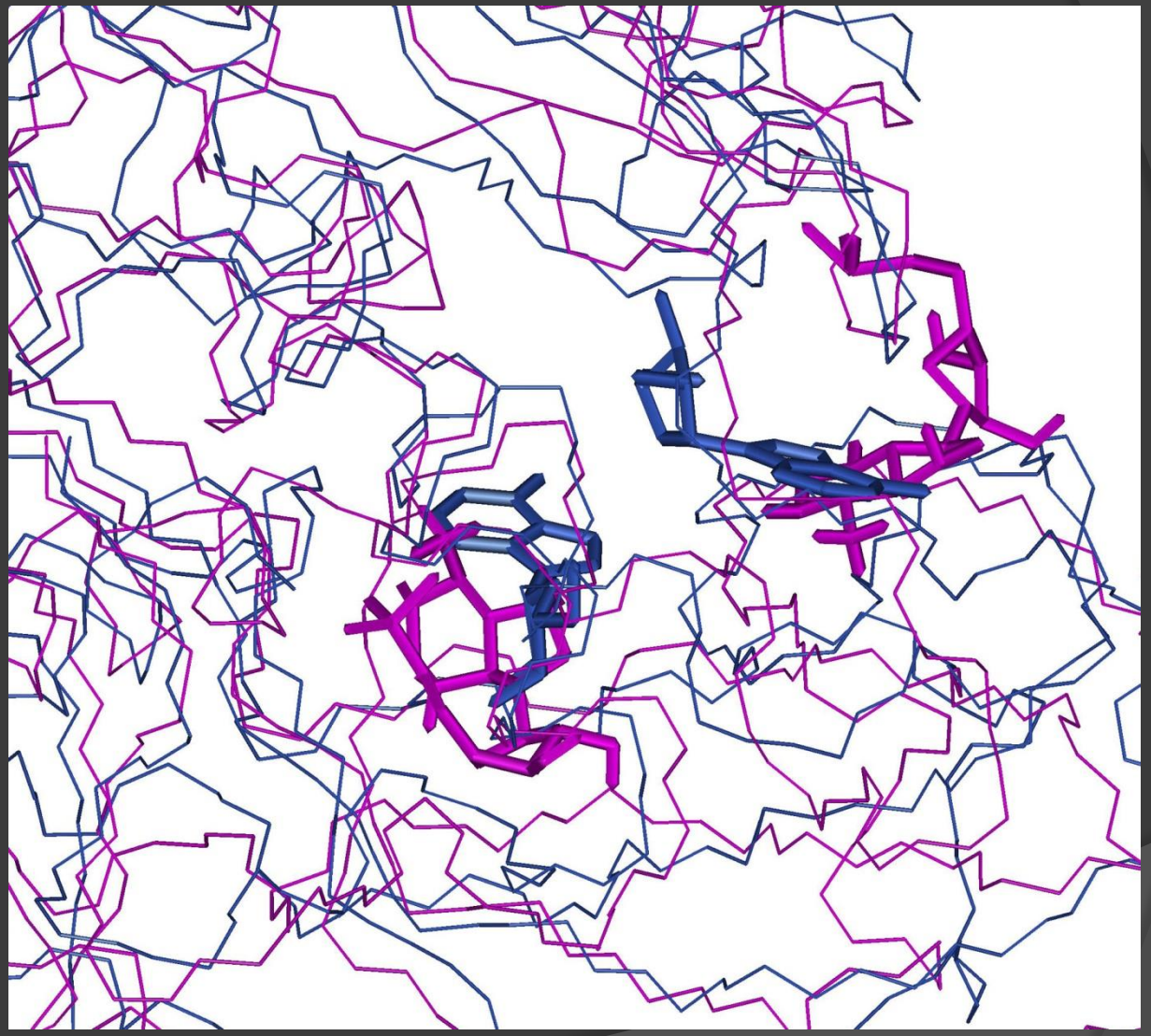
Adenilato Kinasi che da ATP + AMP fornisce 2 ADP

Sistema in cui c'è una miscela delle due proteine ma non c'è ATP viene fornito solo AMP si forma ADP ?

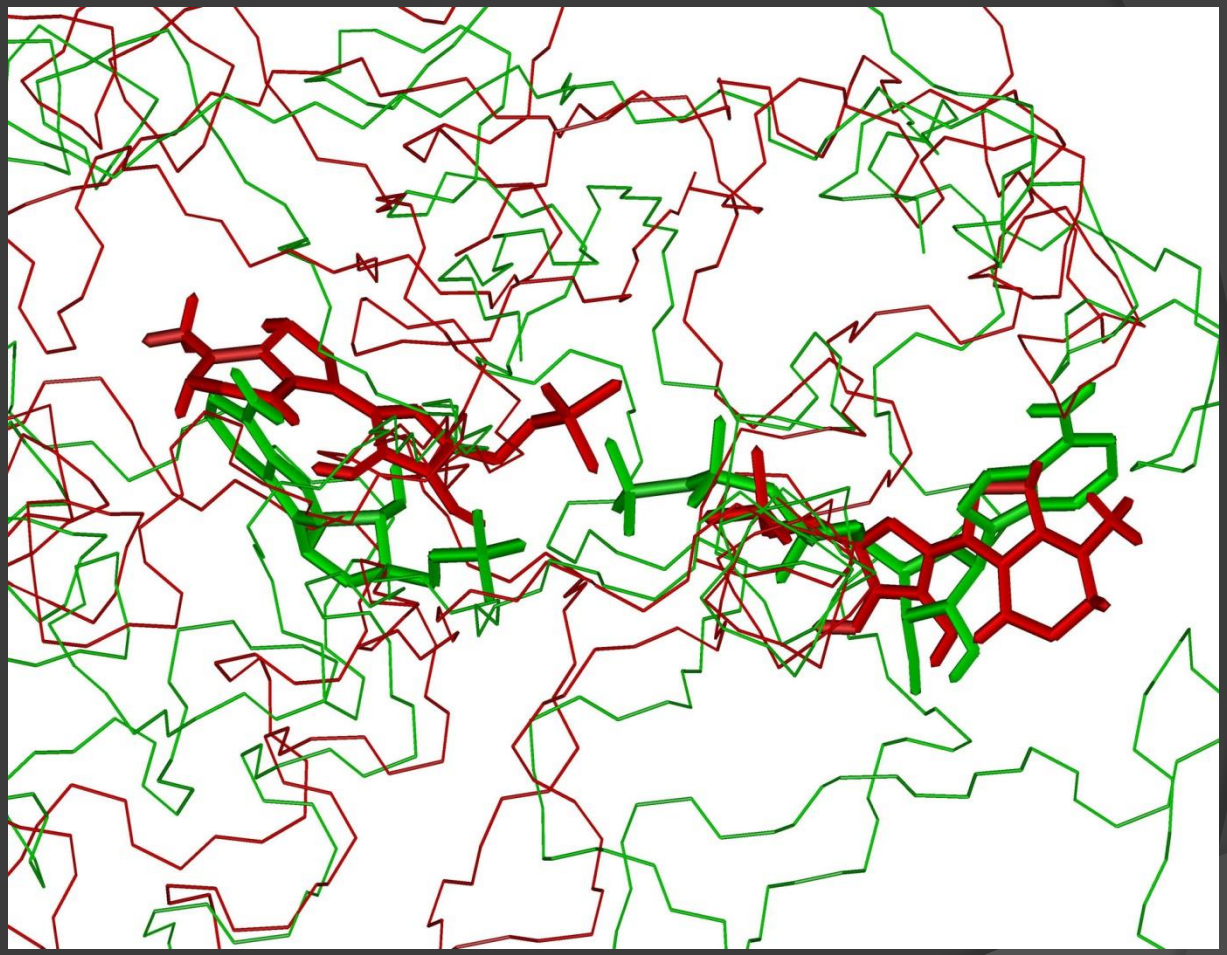
Adenosina Kinasi

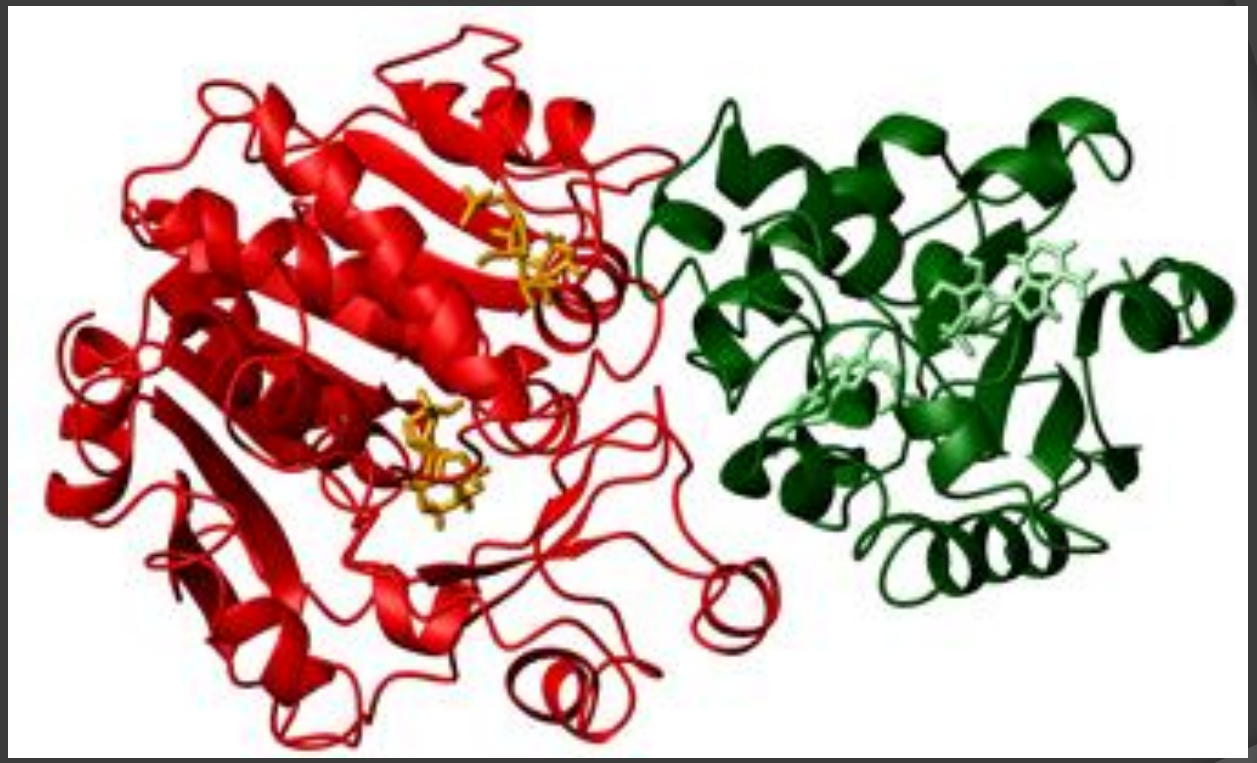


Adenosina
Prima e dopo
la dinamica



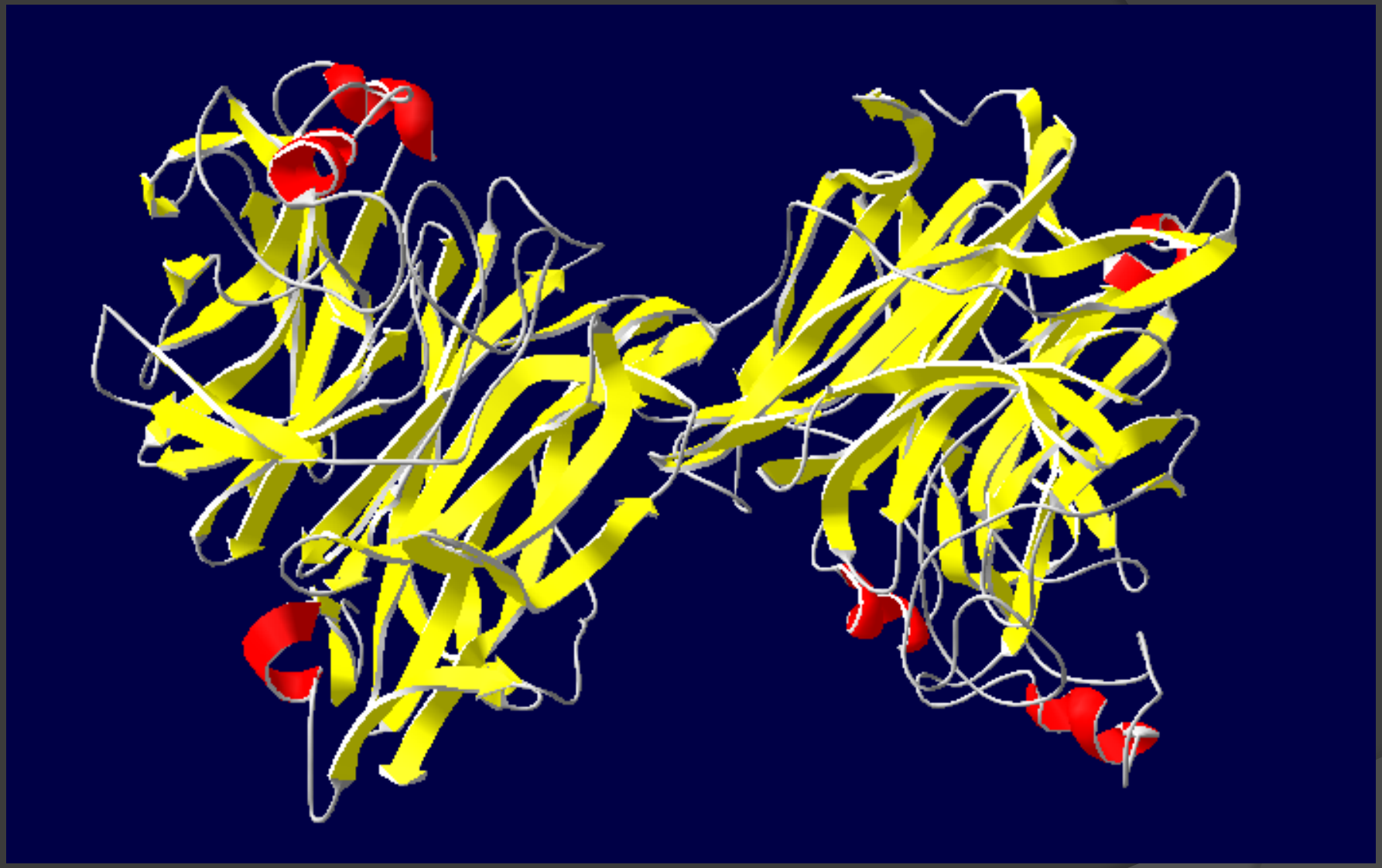
Adenilato
Prima e dopo
la dinamica





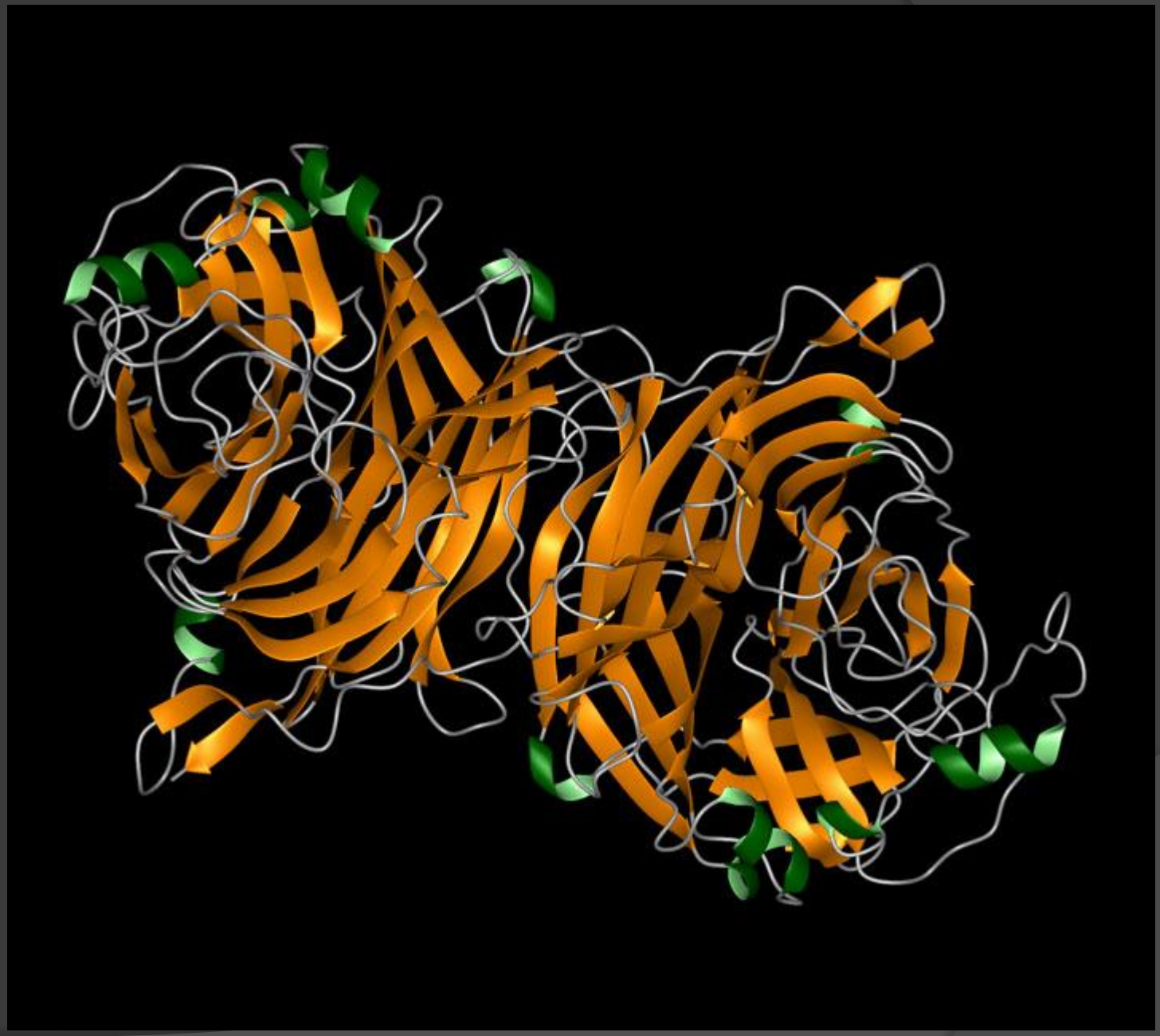
PIVIII

Monomero di HN

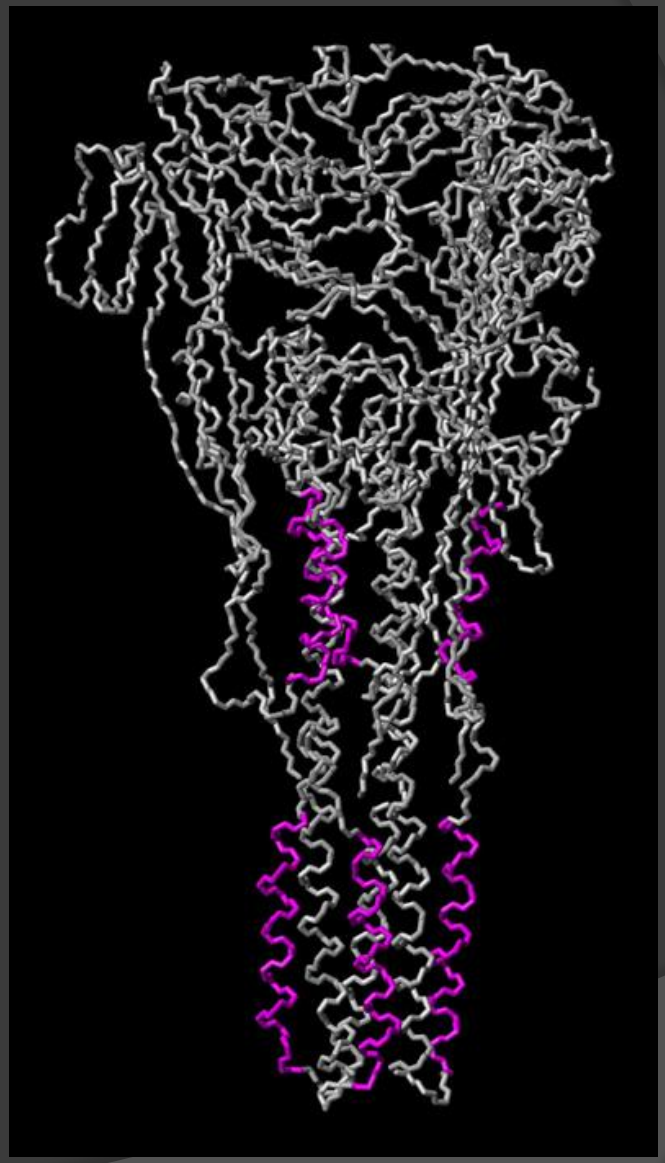


Dimero di HN prima del legame con il recettore cellulare

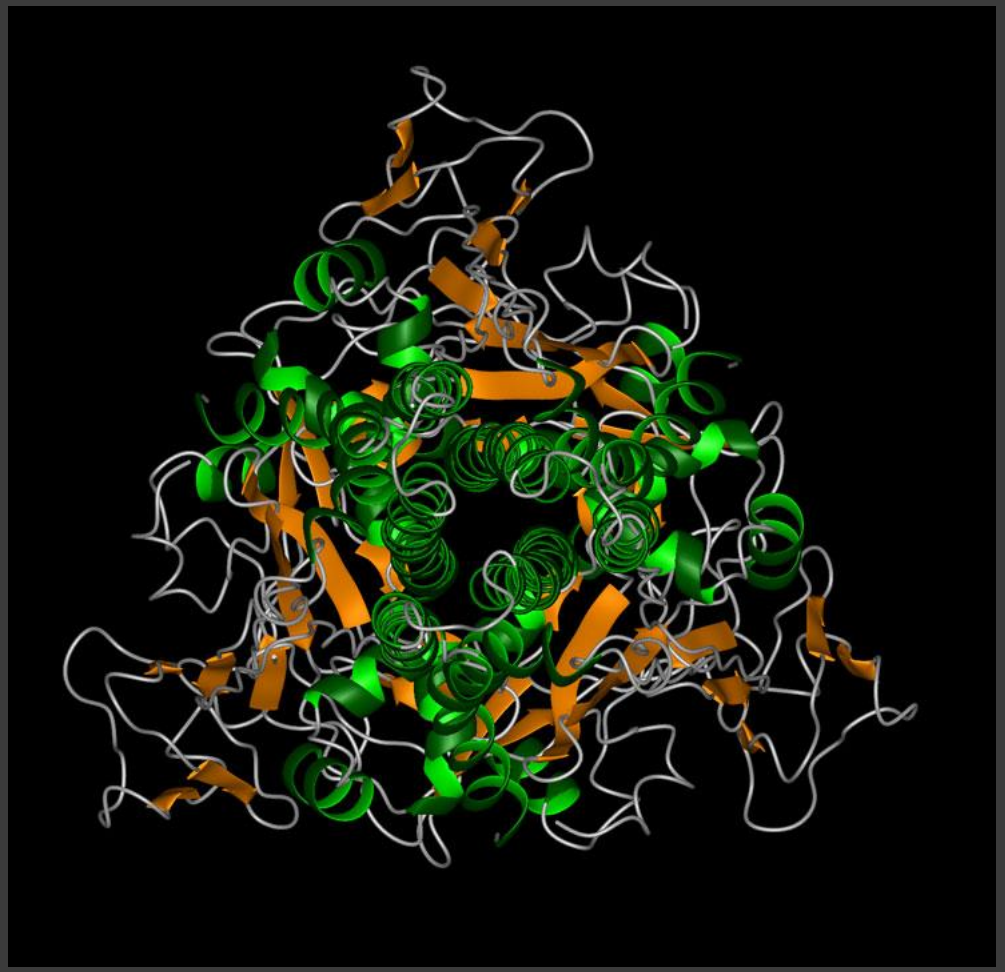
Dimero di HN dopo del legame con il recettore cellulare

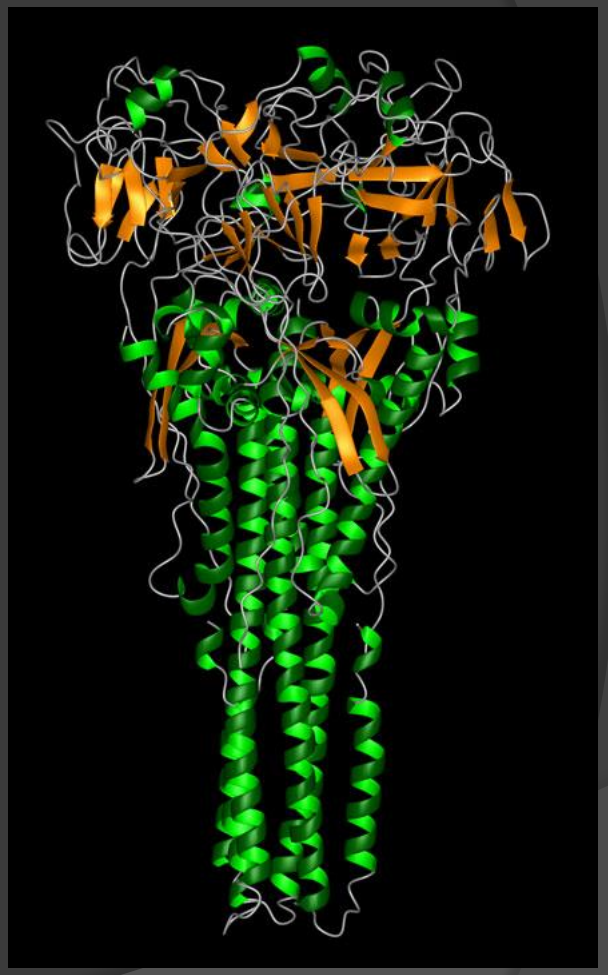


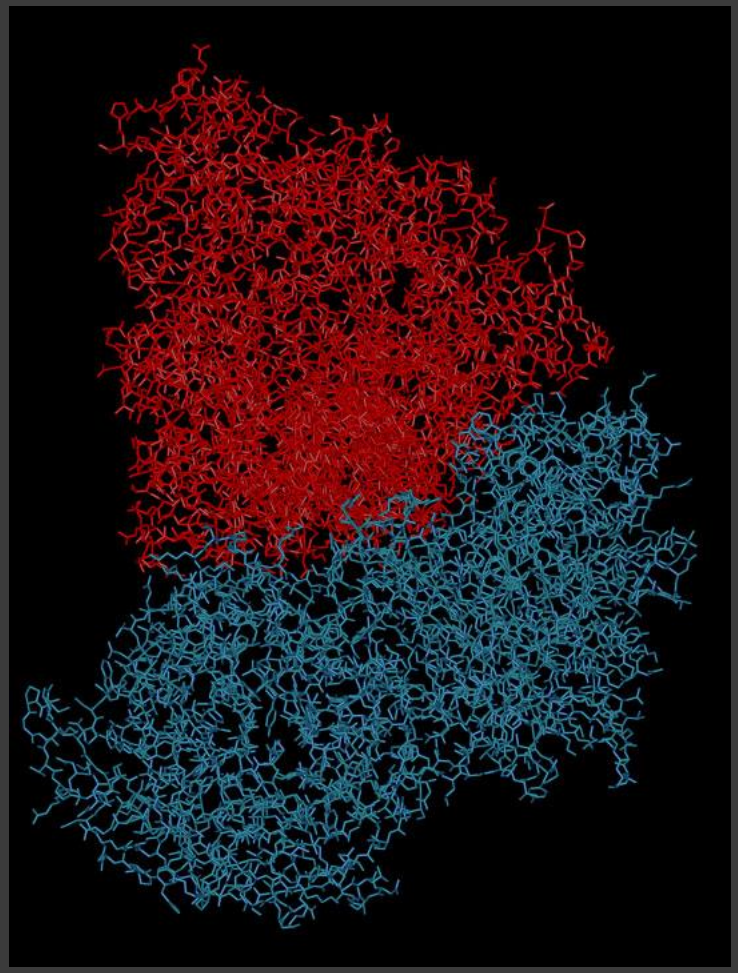
Costruzione del trimero della proteina F

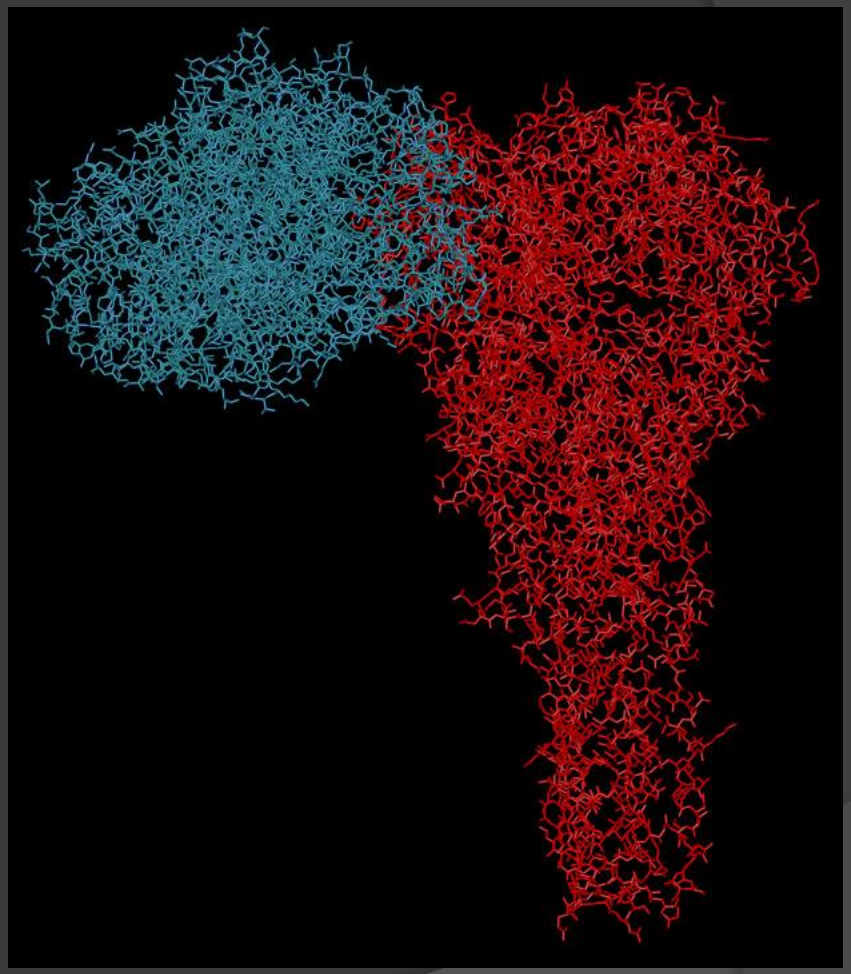


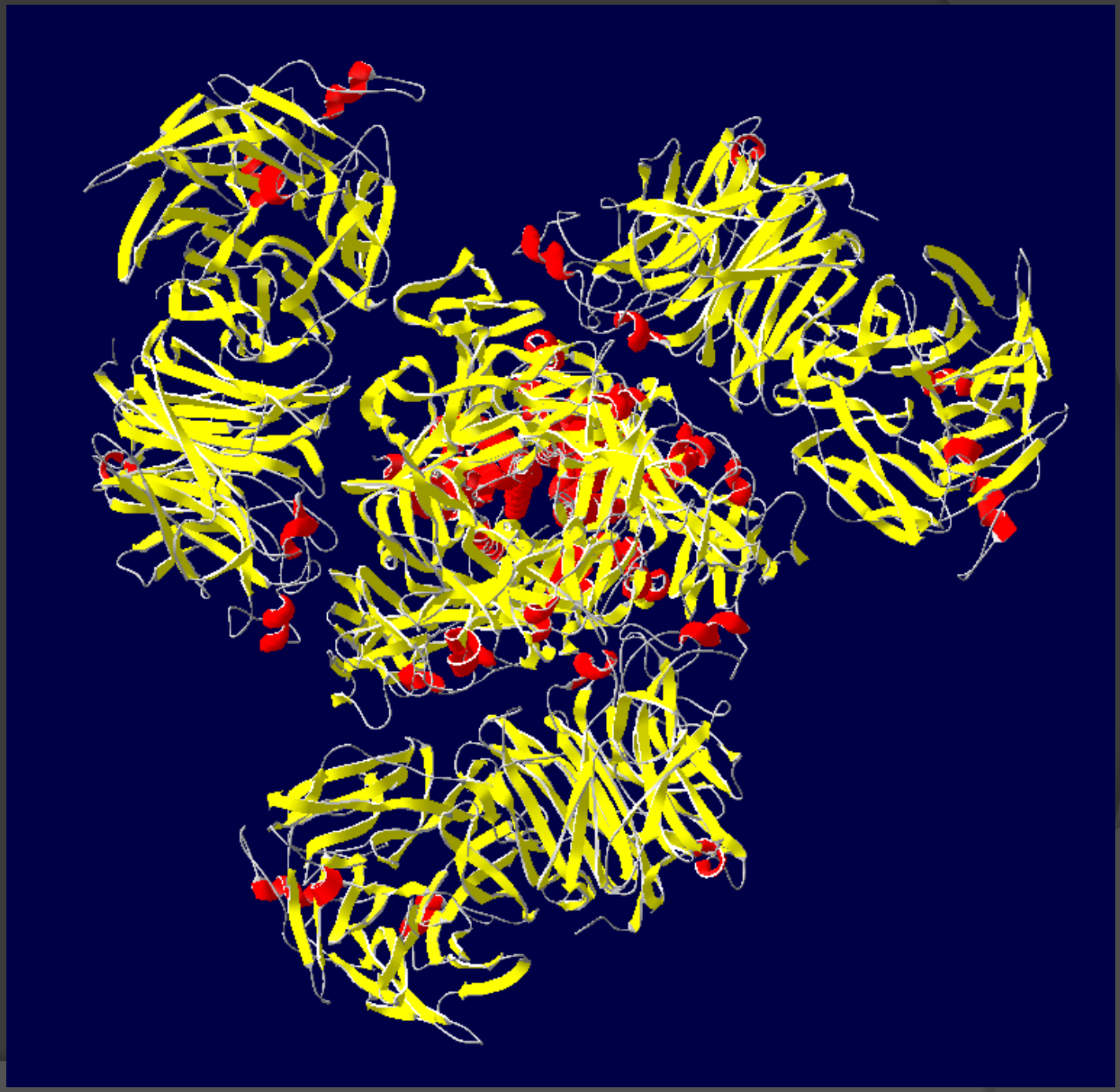
Trimero della proteina F di fusione











INTEGRATED PROJECT

MUCOSAL VACCINES FOR POVERTY-RELATED DISEASES

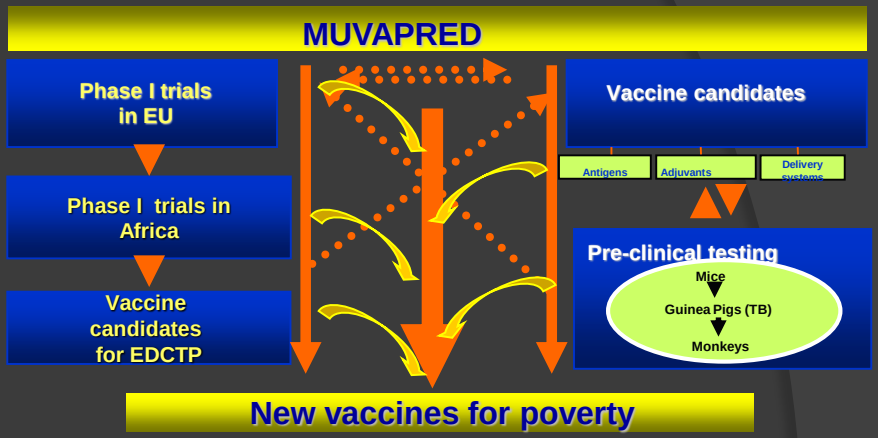
MUVAPRED

Project coordinator: R. Rappuoli, Chiron Srl EC contribution : 15.250.000 Euro
Starting date: 1st December 2003
Duration: 5 years
Participants: 24 from 10 countries

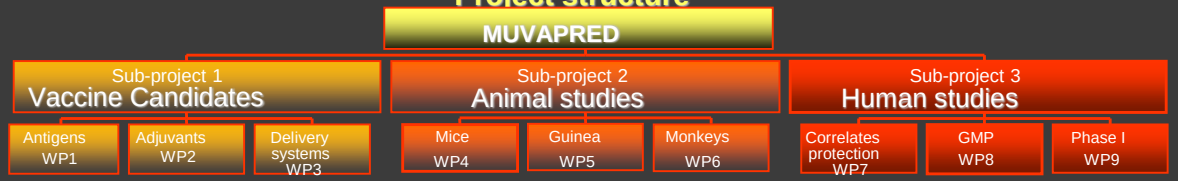
Project Summary

Human Immunodeficiency Virus and *Mycobacterium tuberculosis* enter the human body at mucosal sites. The aim of the present project is to develop mucosally delivered vaccines against HIV and TB which will induce local immunity able to neutralise the pathogens at their port of entry and systemic immunity able to prevent systemic spread of the infection. The possible development of mucosal vaccines against malaria will be also investigated. The trust of the project derives from the recent proof of concept that mucosal vaccines are feasible in humans.

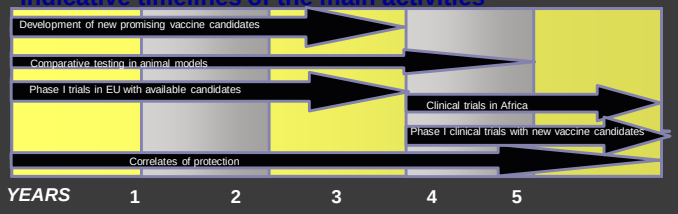
While the first trials are performed, new systems to deliver mucosal vaccines and basic mechanisms of mucosal immune responses and memory in humans will be studied. This will allow better understanding of the clinical results and optimisation of second generation vaccines to be tested in Developing Countries during the second phase of the project.



Project structure



Indicative timelines of the main activities



Objectives

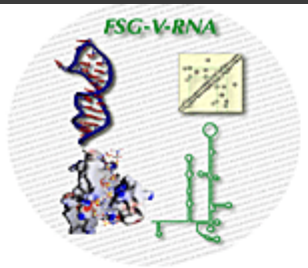
- **Phase I clinical trials in Europe** with the available mucosal antigens against HIV/AIDS and TB (two to three vaccine candidates).
- **Second generation, optimised vaccine candidates** against HIV/AIDS, TB and malaria (antigens, adjuvants, delivery systems).
- Selection of the **most promising vaccine candidates** by comparative testing in animal models
- **Phase I clinical trials in Europe with new vaccine candidates** developed by the consortium
- **Phase I clinical trials in Africa** with the vaccine candidates that have proven to be safe and immunogenic in the first clinical trials in Europe

Scientific and Administrative Management

Participants

Chiron Italy, Imperial College of Science UK, University of Goteborg Sweden, University of Siena Italy, Institute Pasteur France, Istituto Superiore di Sanità Italy, Institute for Research in Biomedicine CH, Max-Planck-Institute, Germany University Hamburg Germany,Trinity College Dublin Ireland, Serum State Institute Denmark, Consiglio Nazionale delle Ricerche Italy, Istituto Humanitas Milano Italy, German Arthritis Research Center DRFZ Germany, St George's Vaccine Institute UK, Institute of Microbiology Czech Republic, CAMR UK, University of Florence Italy, Centre-Hospitalier-Universitaire *Ignace Deen* Guinea, University of Lausanne Switzerland, ALTA Srl Italy, LIONEX Diagnostic & Therapeutics GmbH Germany, INOTECH AG Switzerland

web site: http://www.mucosalimmunity.org/muvapred/

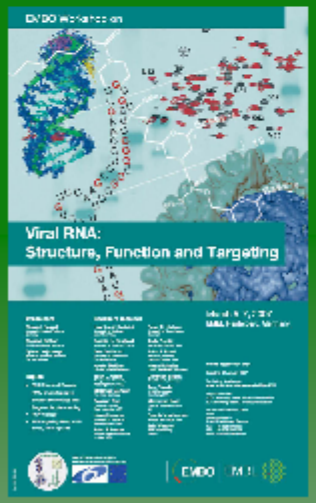


Functional and Structural Genomics of Viral RNA

6th Framework Program EU

Login:
Pass:
Go

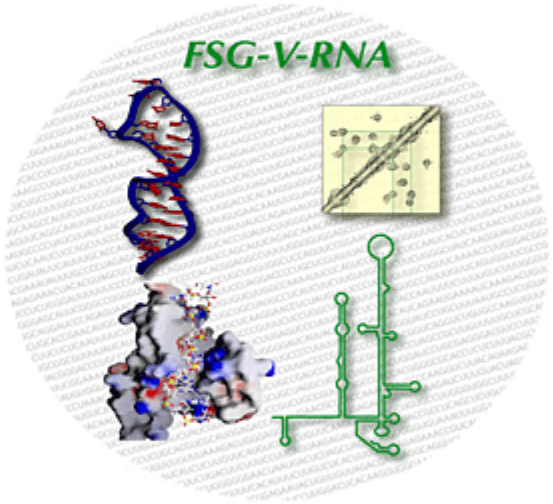
Don't miss the
FSGVRNA WORKSHOP 2007
(Heidelberg, Germany)
5-7 March 2007



The current project joins, in an interdisciplinary fashion, leading European labs to integrate their equipment and expertises on:

» The structural, functional and virological analysis of RNA and RNA-protein complexes from viruses.

» The evaluation of these viral RNAs as targets for novel types of drugs, either small RNA binding compounds, or antiviral RNAs.



Acronym: FSG-V-RNA

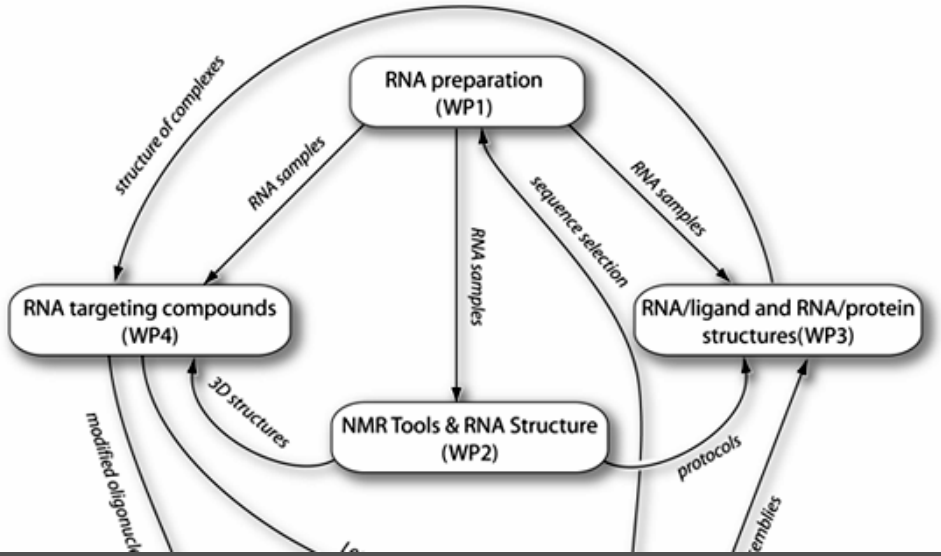
Research Topic: Life sciences, genomics and biotechnology for health, LSH-2002-1.1.0-1: For STREP and CA, research should focus on multidisciplinary functional genomics approaches. Proposals concerned with the development of new tools and approaches, including the standardisation of protocols, to facilitate generation of new knowledge in functional and structural genomics are also envisaged.

The STREP will develop and improve tools and approaches to facilitate the generation of new knowledge in functional and structural genomics of viral RNAs. Specifically, (i) new methods and tools for the rapid and efficient structural analysis of RNA and RNA-protein complexes will be developed, (ii) these optimised tools will be applied to essential RNA elements that are vital for the function of HBV, HCV and HIV viruses, and (iii) complementary screening techniques and structure analysis of RNA-ligand complexes will be implemented to promote the identification of antiviral compounds targeting these RNA structures. The project exploits available RNA sequence data but will also expand our knowledge of viral RNA sequence elements and their variations.

The biomedical importance of RNA as a research target is stressed by the fact that viral infections, such as HBV, HCV and HIV are major global public health problems. The outcomes of the project are expected to initiate the development of novel drugs that target viral RNA molecules and have thus strong implications for public health. The research consortium is strengthened by the involvement of an SME, which contributes a proprietary RNA screening method. This SME will also function as a stakeholder to support the downstream application development towards RNA targeting drugs - even though this is beyond the scope of the current project.

The project involves multidisciplinary research which is only possible by integrating, in an interdisciplinary fashion, the research capacities of a number of leading European labs, i.e. their equipment and complementary expertise on the structural, functional and virological analysis of RNA and RNA-ligand complexes. The innovative tools developed will be made available to other researchers throughout Europe and open the way for efficient analysis of a wide range of RNA-based processes extending far beyond analysis of viral RNAs.

Work packages



3D repertoire

Funded by EU Framework 6

Jobs | Contact

Home | Complexes | Press & Publications | Workshops & Meetings | Methods | Structural Genomics | Private

Project | Summary | Objectives | Workpackages | Partners

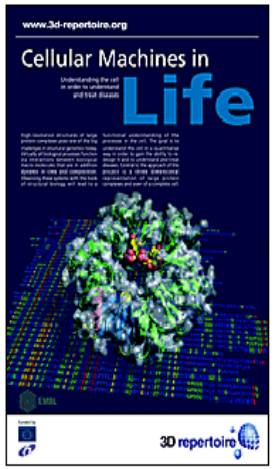
Welcome

In the last two decades biological science has made huge progress in many areas. We live in an era when the entire genetic code of many organisms has been established. However, to make full use of this treasure of information, we need to bring it together with the knowledge of what the products of all these genes, the proteins, are doing. At present, even though we appear to have a vast amount of information at our disposal, it would be unreasonable to expect that in the near or medium term we could reach this level of understanding for an entire cell. So, a pre-requisite for this goal is comprehensive knowledge of the biological functions of the complete set of genes and proteins within a genome (post-genomic biology).

Introduction

Proteins rarely act alone: they typically interact with other macromolecules to perform particular cellular tasks. The resulting functional assemblies (complexes) achieve more than the sum of their parts and these complexes have functions that cannot easily understood by even the most systematic analyses of individual proteins. So, the discovery and analysis of particular cellular protein complexes under physiological conditions provides key insights into their function and takes characterisation of cellular systems well beyond the limits of other experiments. Prominent examples include the ribosome, the chaperonin GroEl/GroEs, the spliceosome, the cyclosome, the proteasome, the nuclear pore complex and the synaptosome. Analyses of results from genome-scale interaction experiments in yeast show a clear tendency for many yeast assemblies to mirror their equivalents in animals, including the model organisms and man. Complexes essential for normal cell activitiy overlap significantly and represent the building blocks of a Eukaryotic core proteome, covering basic cellular function. More importantly, those conserved between yeast and man contribute significantly to the understanding of multifactor diseases, particularly those related to key cellular processes. Elucidation of three-dimensional (3D) structures for protein complexes will open new avenues to unravel the molecular pathology and physiology of human diseases, leading to rational, target-oriented therapeutic approaches. Moreover developments in 3D tomography show that it will soon be possible to fit such structures into a whole cell tomogram. This will be a great leap for systems biology, since it will place complexes in their precise cellular context and provide critical concentration information essential for the quantitative understanding of a living cell. However, without the individual complexes, it will be exceedingly difficult to understand such whole cell images.

3D-Repertoire is an integrated project funded by the European Commission under Framework 6. It brings together scientists from 14 institutions and private companies, in seven countries across Europe, with the aim to resolve structures for all amenable protein complexes from budding yeast (or where necessary equivalents from other species) at the best possible resolution by electron microscopy, X-ray crystallography and in silico approximations. The team of top scientists formed through 3D-Repertoire covers the wide range of expertise needed for such an ambitious undertaking and we have secured access to the yeast constructions used to do the complete pull-down study of complexes in Yeast (given to Euroscarf by Cellzome).



News

2nd Annual Meeting, 22-23/02/07

3D repertoire

2nd 3D Repertoire Annual Meeting