Bioinformatics

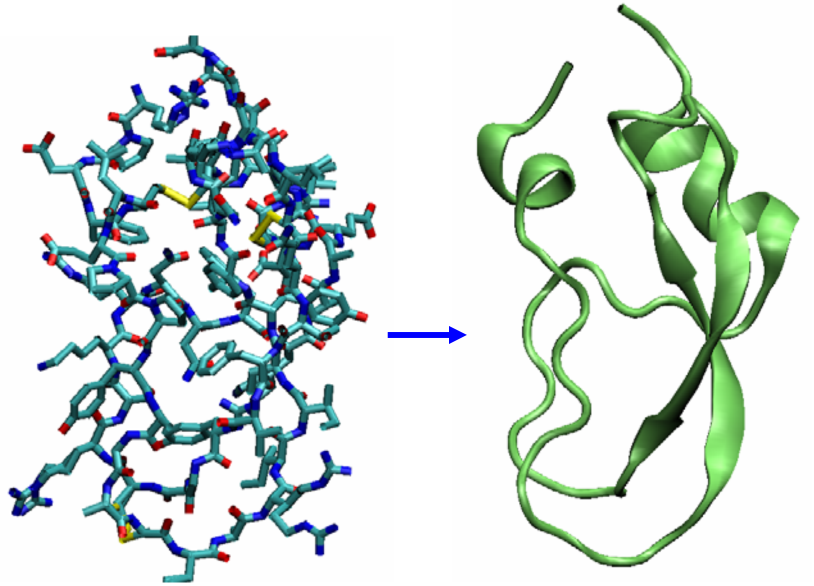
Use of computer to analyze and archive biological data (sequence and structural information) on a large scale

- includes development of analysis algorithms, visualization software, database design

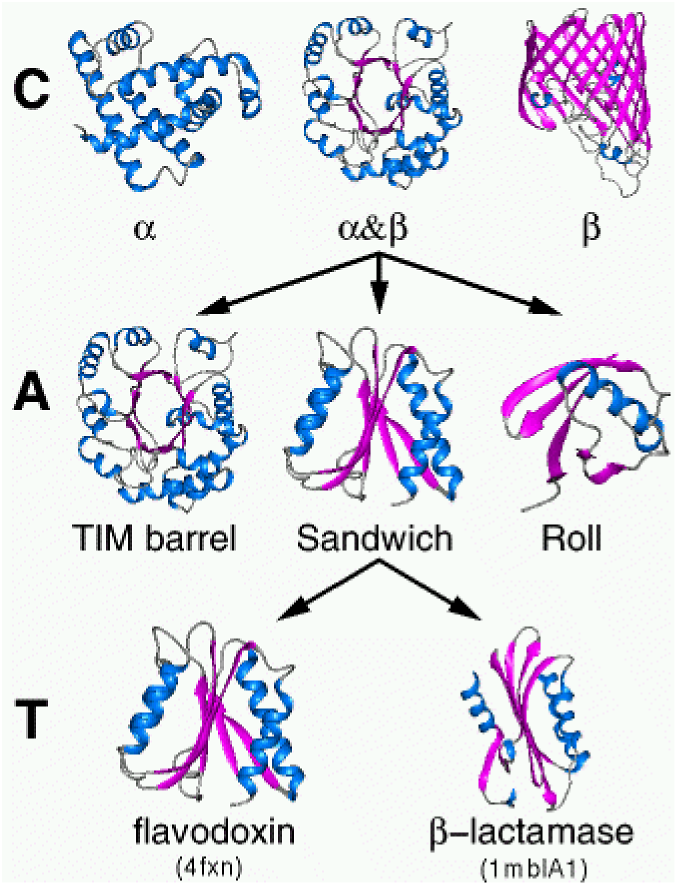
- Secondary structure assignment
- Secondary structure prediction
- Sequence alignment
- Structural alignment
- Tertiary structure prediction

Secondary structure assignment

Easy visualization



Structural classification

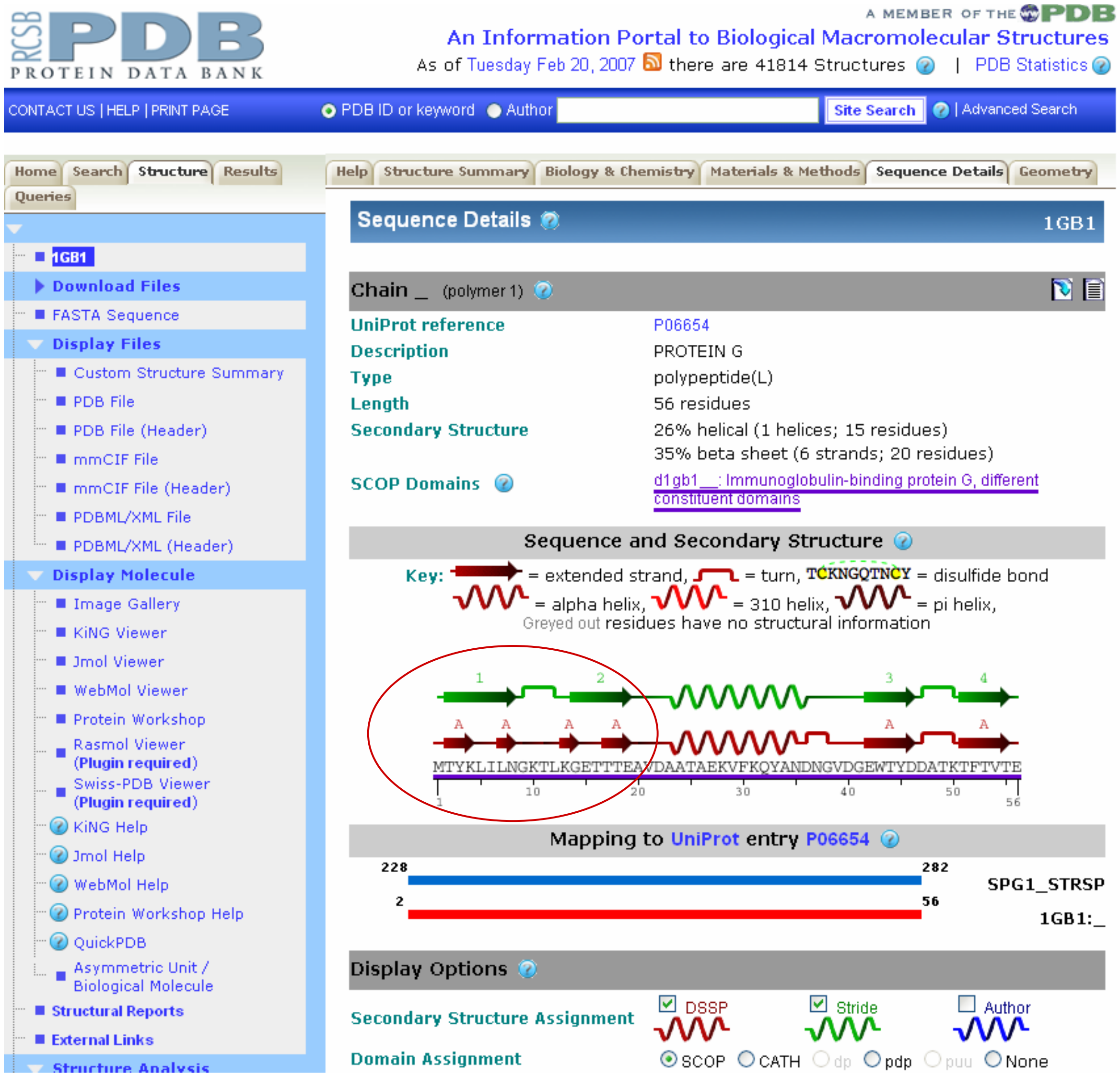


Detection of structural motifs and improved sequence-structure searches

Structural alignment

Given a structure, identify the regions of secondary structure

- DSSP, Stride, DEFINE
- implementation dependent



RCSB PDB PROTEIN DATA BANK

A MEMBER OF THE PDB

An Information Portal to Biological Macromolecular Structures

As of Tuesday Feb 20, 2007 there are 41814 Structures | PDB Statistics

CONTACT US | HELP | PRINT PAGE ● PDB ID or keyword ○ Author Site Search | Advanced Search

Home | Search | Structure | Results | Queries

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

- 1GB1
- Download Files
- FASTA Sequence
- Display Files
  - Custom Structure Summary
  - PDB File
  - PDB File (Header)
  - mmCIF File
  - mmCIF File (Header)
  - PDBML/XML File
  - PDBML/XML (Header)
- Display Molecule
  - Image Gallery
  - KiNG Viewer
  - Jmol Viewer
  - WebMol Viewer
  - Protein Workshop
  - Rasmol Viewer (Plugin required)
  - Swiss-PDB Viewer (Plugin required)
  - KiNG Help
  - Jmol Help
  - WebMol Help
  - Protein Workshop Help
  - QuickPDB
  - Asymmetric Unit / Biological Molecule
- Structural Reports
- External Links
- Structure Analysis

Sequence Details 1GB1

Chain _ (polymer 1)

UniProt reference	P06654
Description	PROTEIN G
Type	polypeptide(L)
Length	56 residues
Secondary Structure	26% helical (1 helices; 15 residues) 35% beta sheet (6 strands; 20 residues)
SCOP Domains	d1gb1__: Immunoglobulin-binding protein G, different constituent domains

Sequence and Secondary Structure

Key: ➡ = extended strand, ⎍ = turn, TCKNGQTNCY = disulfide bond
∿ = alpha helix, ∿ = 310 helix, ∿ = pi helix,
Greyed out residues have no structural information

MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE

Mapping to UniProt entry P06654

Display Options

Secondary Structure Assignment ☑ DSSP ☑ Stride □ Author

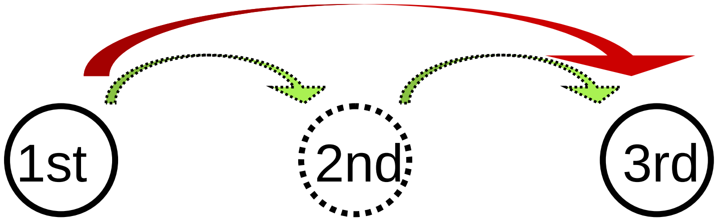
Domain Assignment ● SCOP ○ CATH ○ dp ○ pdp ○ puu ○ None

Secondary structure prediction

Tertiary structure prediction from the amino acid sequence is very difficult

Prediction of secondary structure is feasible and more reliable

In some models of protein folding, secondary structural elements form first before a tertiary structure is formed



A-C-H-Y-T-T-E-K-R-G-G-S-G-T-K-K-R-E-A

↓

H-H-H-H-H-H-H-H-C-C-C-C-C-S-S-S-S-S-S

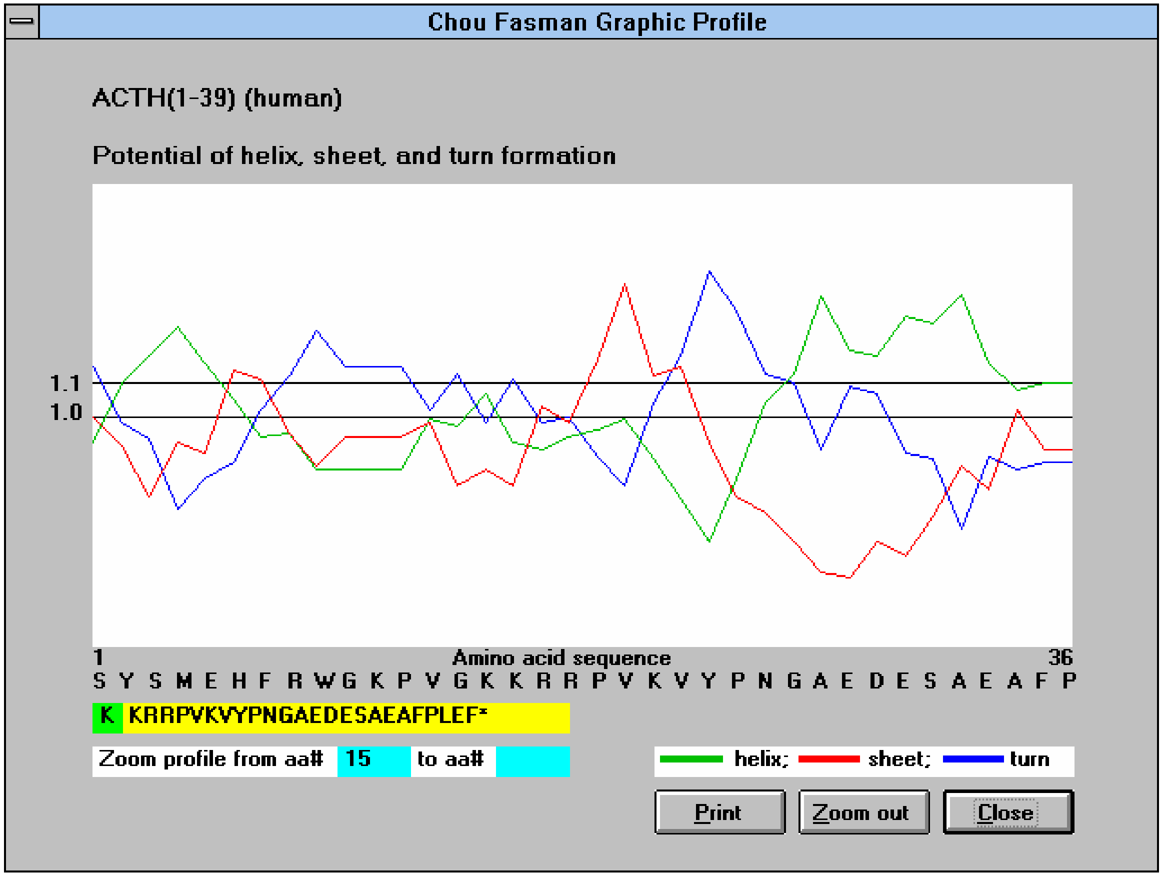
Knowing the region of secondary structure is critical for some applications

- transmembrane domain of the membrane protein GPCR
- secondary structural info may be sufficient for some studies

Prediction methods

Use known secondary structure propensities of individual amino acids—either statistical or experimental

- helix former, helix breaker, helix neutral, sheet former, sheet breaker, etc
- develop heuristic rules for identifying and extending a helix or a sheet



Chou Fasman Graphic Profile

ACTH(1-39) (human)

Potential of helix, sheet, and turn formation

K KRRPVKVYPNGAEDESAEAFPLEF*

Zoom profile from aa# 15 to aa#

1.1
1.0
1
Amino acid sequence
36
S Y S M E H F R W G K P V G K K R R P V K V Y P N G A E D E S A E A F P
helix; sheet; turn
Print
Zoom out
Close

Examine sets of adjacent amino acids (e.g. windows of 11-21 amino acids) rather than individual amino acids

- probability of an amino acid to be in a particular secondary structure considering the nearby residues
- HHSLCSHSHHSC less likely than HHHHHCCCSSSS
- local context is important

Secondary structure prediction services

- PredictProtein
- PHD
- JPRED
- PSSP

Limitations

- overall prediction: 60%
- beta-strands prediction: ~35%
- predictions include small secondary elements that cannot be easily integrated into longer structures

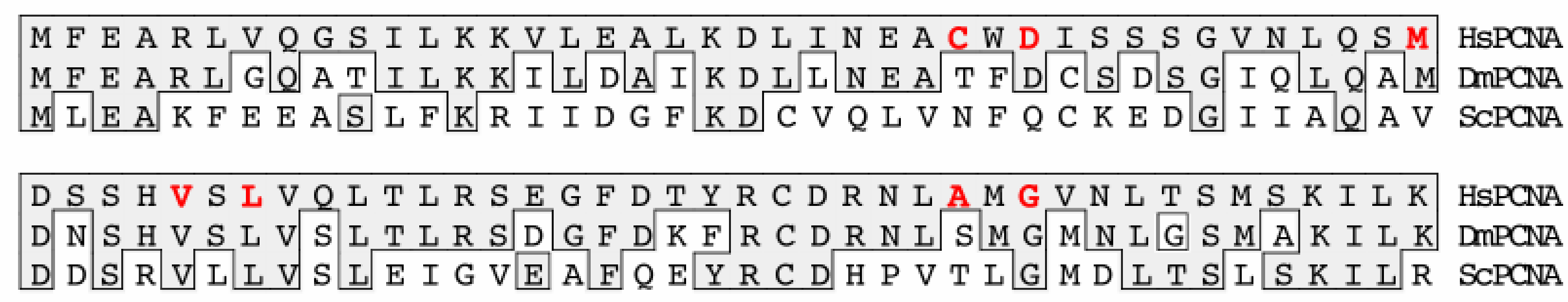
Sequence alignment

Process of comparing two or more sequences by looking for a series of individual characters or character patterns (similar vs. identical) that are in the same order in the sequences

Sequence alignment lies at the heart of bioinformatics

- newly discovered sequence may be related to known sequence
- models evolutionary relationship
- assist in engineering and 3D prediction
- basis to functional genomics
- population genomics—genetic variations in an isolated group (DeCode).

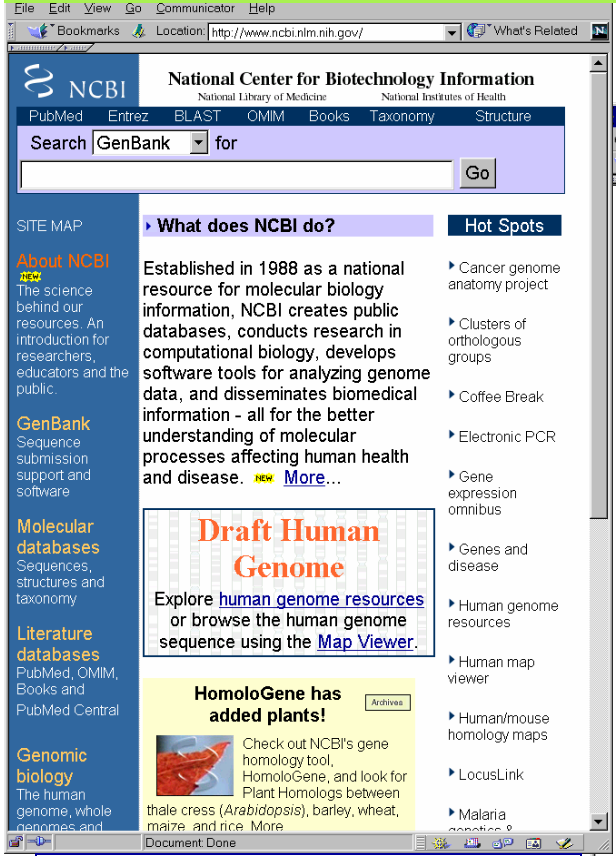
Identity vs. similarity – definition of similarity



```
M F E A R L V Q G S I L K K V L E A L K D L I N E A C W D I S S S G V N L Q S M  HsPCNA
M F E A R L G Q A T I L K K I L D A I K D L L N E A T F D C S D S G I Q L Q A M  DmPCNA
M L E A K F E E A S L F K R I I D G F K D C V Q L V N F Q C K E D G I I A Q A V  ScPCNA

D S S H V S L V Q L T L R S E G F D T Y R C D R N L A M G V N L T S M S K I L K  HsPCNA
D N S H V S L V S L T L R S D G F D K F R C D R N L S M G M N L G S M A K I L K  DmPCNA
D D S R V L L V S L E I G V E A F Q E Y R C D H P V T L G M D L T S L S K I L R  ScPCNA
```

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


File Edit View Go Communicator Help
Bookmarks
Location: http://www.ncbi.nlm.nih.gov/
What's Related
NCBI
PubMed Entrez BLAST OMIM Books Taxonomy Structure
Search GenBank for
Go
Archives
Document: Done

National Center for Biotechnology Information

National Library of Medicine National Institutes of Health

SITE MAP

About NCBI NEW The science behind our resources. An introduction for researchers, educators and the public.

GenBank Sequence submission support and software

Molecular databases Sequences, structures and taxonomy

Literature databases PubMed, OMIM, Books and PubMed Central

Genomic biology The human genome, whole genomes and

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. NEW More...

Draft Human Genome

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

HomoloGene has added plants!

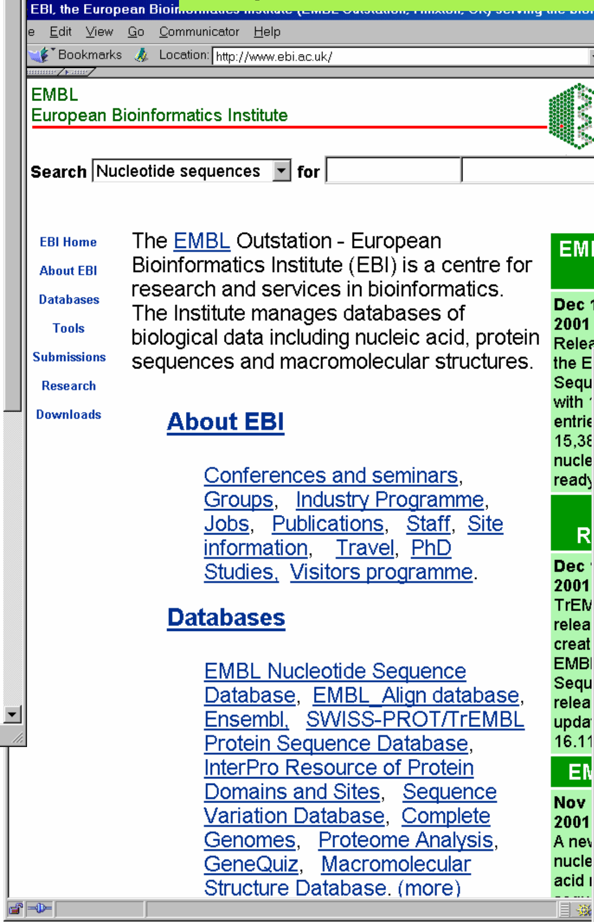
Check out NCBI's gene homology tool, HomoloGene, and look for Plant Homologs between thale cress (*Arabidopsis*), barley, wheat, maize, and rice. More

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
- LocusLink
- Malaria genetics &

National **C**enter for
Biotechnology
Information
National **L**ibrary of
Medicine
National **I**nstitute of
Heatlth
USA

http://www.ebi.ac.uk/


EBI, the European Bioi... - N...
e Edit View Go Communicator Help
Bookmarks
Location: http://www.ebi.ac.uk/
What's Related
Search Nucleotide sequences for
Go ?

EMBL
European Bioinformatics Institute

EBI Home
About EBI
Databases
Tools
Submissions
Research
Downloads

The EMBL Outstation - European Bioinformatics Institute (EBI) is a centre for research and services in bioinformatics. The Institute manages databases of biological data including nucleic acid, protein sequences and macromolecular structures.

About EBI

Conferences and seminars, Groups, Industry Programme, Jobs, Publications, Staff, Site information, Travel, PhD Studies, Visitors programme.

Databases

EMBL Nucleotide Sequence Database, EMBL_Align database, Ensembl, SWISS-PROT/TrEMBL Protein Sequence Database, InterPro Resource of Protein Domains and Sites, Sequence Variation Database, Complete Genomes, Proteome Analysis, GeneQuiz, Macromolecular Structure Database. (more)

EMBL Release 69

Dec 18th 2001
Release 69 of
the E
Sequ
with
entrie
15,38
nucle
ready

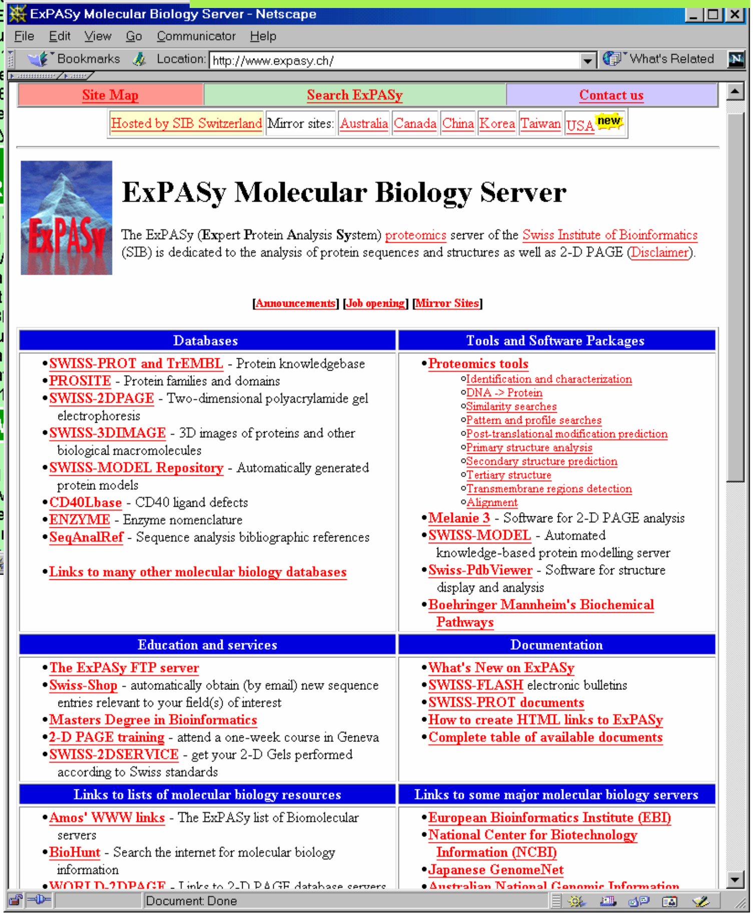
Dec
2001
TrEM
relea
creat
EMBI
Sequ
relea
upda
16.11

Nov
2001
A nev
nucle
acid

Eouropean
Bioinformatics **I**nstitute
European **M**olecular
Biology **L**aboratory
UK

http://www.expasy.ch/

Expert **P**rotein **A**nalysis
System proteomics
server of the Swiss
Institute of
Bioinformatics


ExPASy Molecular Biology Server - Netscape
File Edit View Go Communicator Help
Bookmarks
Location: http://www.expasy.ch/
What's Related
ExPASy
Document: Done

Site Map | Search ExPASy | Contact us

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

ExPASy Molecular Biology 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).

[Announcements] [Job opening] [Mirror Sites]

Databases

- SWISS-PROT and TrEMBL - Protein knowledgebase
- PROSITE - Protein families and domains
- SWISS-2DPAGE - Two-dimensional polyacrylamide gel electrophoresis
- SWISS-3DIMAGE - 3D images of proteins and other biological macromolecules
- SWISS-MODEL Repository - Automatically generated protein models
- CD40Lbase - CD40 ligand defects
- ENZYME - Enzyme nomenclature
- SeqAnalRef - Sequence analysis bibliographic references
- Links to many other molecular biology databases

Tools and Software Packages

- Proteomics tools
  - Identification and characterization
  - DNA -> Protein
  - Similarity searches
  - Pattern and profile searches
  - Post-translational modification prediction
  - Primary structure analysis
  - Secondary structure prediction
  - Tertiary structure
  - Transmembrane regions detection
  - Alignment
- Melanie 3 - Software for 2-D PAGE analysis
- SWISS-MODEL - Automated knowledge-based protein modelling server
- Swiss-PdbViewer - Software for structure display and analysis
- Boehringer Mannheim's Biochemical Pathways

Education and services

- The ExPASy FTP server
- Swiss-Shop - automatically obtain (by email) new sequence entries relevant to your field(s) of interest
- Masters Degree in Bioinformatics
- 2-D PAGE training - attend a one-week course in Geneva
- SWISS-2DSERVICE - get your 2-D Gels performed according to Swiss standards

Documentation

- What's New on ExPASy
- SWISS-FLASH electronic bulletins
- SWISS-PROT documents
- How to create HTML links to ExPASy
- Complete table of available documents

Links to lists of molecular biology resources

- Amos' WWW links - The ExPASy list of Biomolecular servers
- BioHunt - Search the internet for molecular biology information
- WORLD-2DPAGE - Links to 2-D PAGE database servers

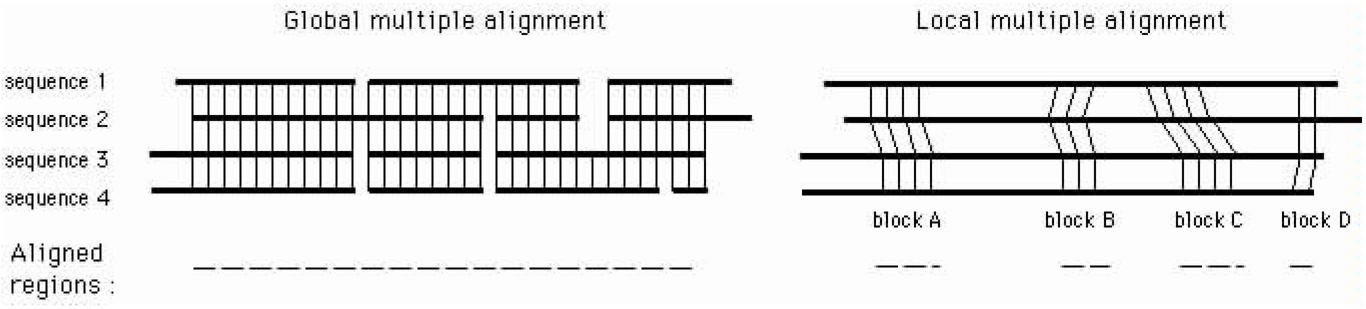
Links to some major molecular biology servers

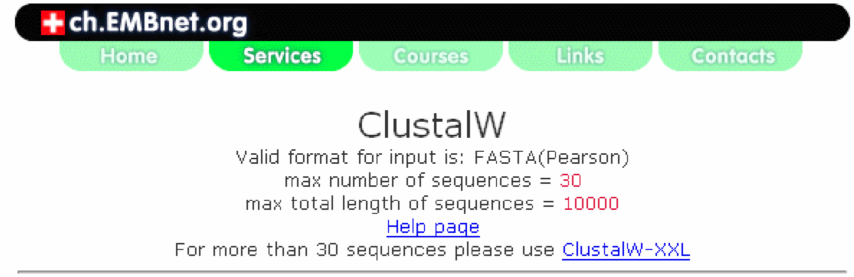
- European Bioinformatics Institute (EBI)
- National Center for Biotechnology Information (NCBI)
- Japanese GenomeNet
- Australian National Genomic Information

Multiple sequence alignment (MSA)

Incorporate evolutionary information through multiple sequence alignment

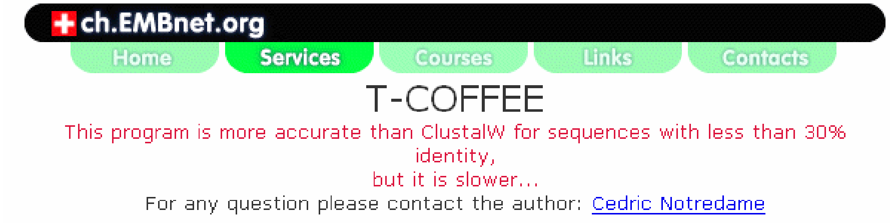
- information on sequence conservation, substitution, and potential interaction
- ClustalW
- T-Coffee





ClustalW

Valid format for input is: FASTA(Pearson)
max number of sequences = 30
max total length of sequences = 10000
Help page
For more than 30 sequences please use ClustalW-XXL



T-COFFEE

This program is more accurate than ClustalW for sequences with less than 30% identity,
but it is slower...

For any question please contact the author: Cedric Notredame

Structural alignment

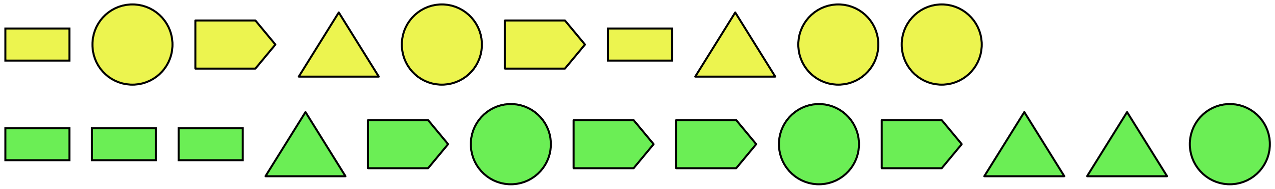
Structures are more conserved than sequences

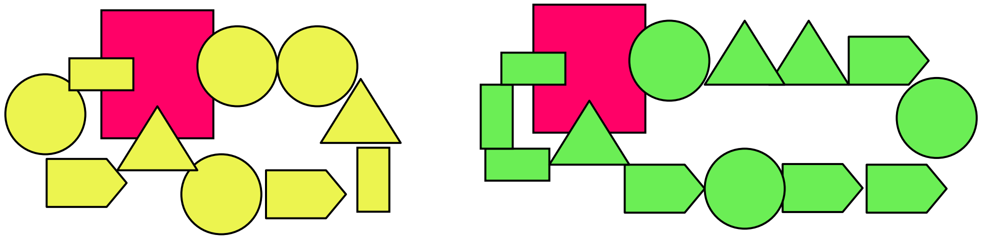
In the “twilight zone” of sequence similarity, structural alignment might help to correctly determine the relations between two proteins

Structural alignment is more predictive of function than sequence alignment

sequence 1

sequence 2





similar local structure

Alignment v. superposition

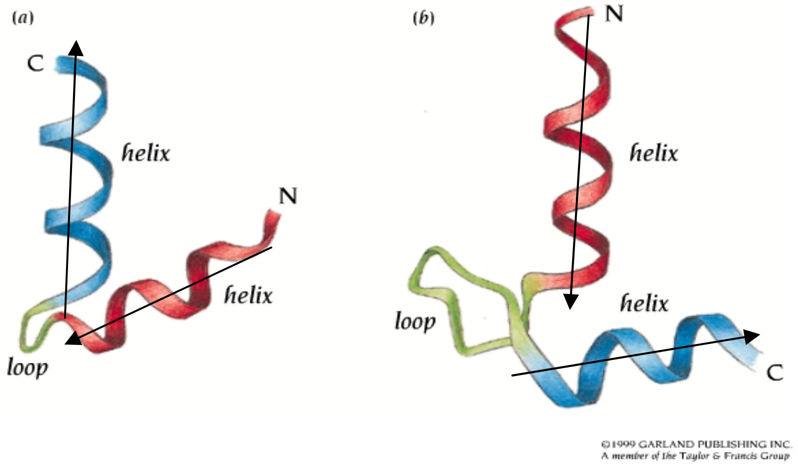
Superposition assumes the two are related—translate and rotate one of them to minimize the total rmsd

Alignment is a means of determining if two are structurally related by mapping stretches of atoms from one protein to another

- integral to structural classification

- Distance alignment matrix (DALI)
- Combinatorial extension (CE)
- Sequential Structure Alignment Program (SSAP)
- Spatial Arrangements of Backbone Fragments (SARF2)
- Structural Alignment of Multiple Proteins (STAMP)
- Structure based Alignment Program (STRAP)

Many are available as web services

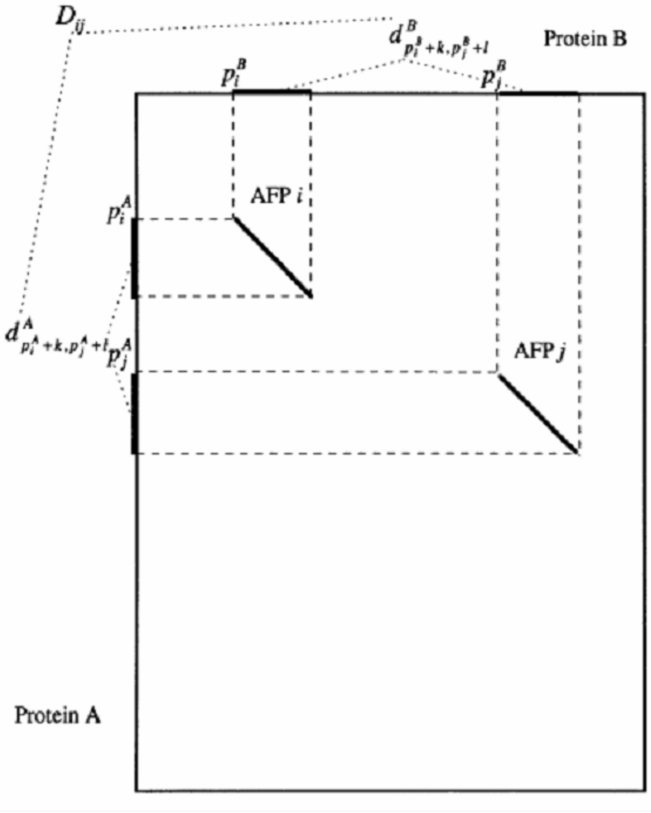


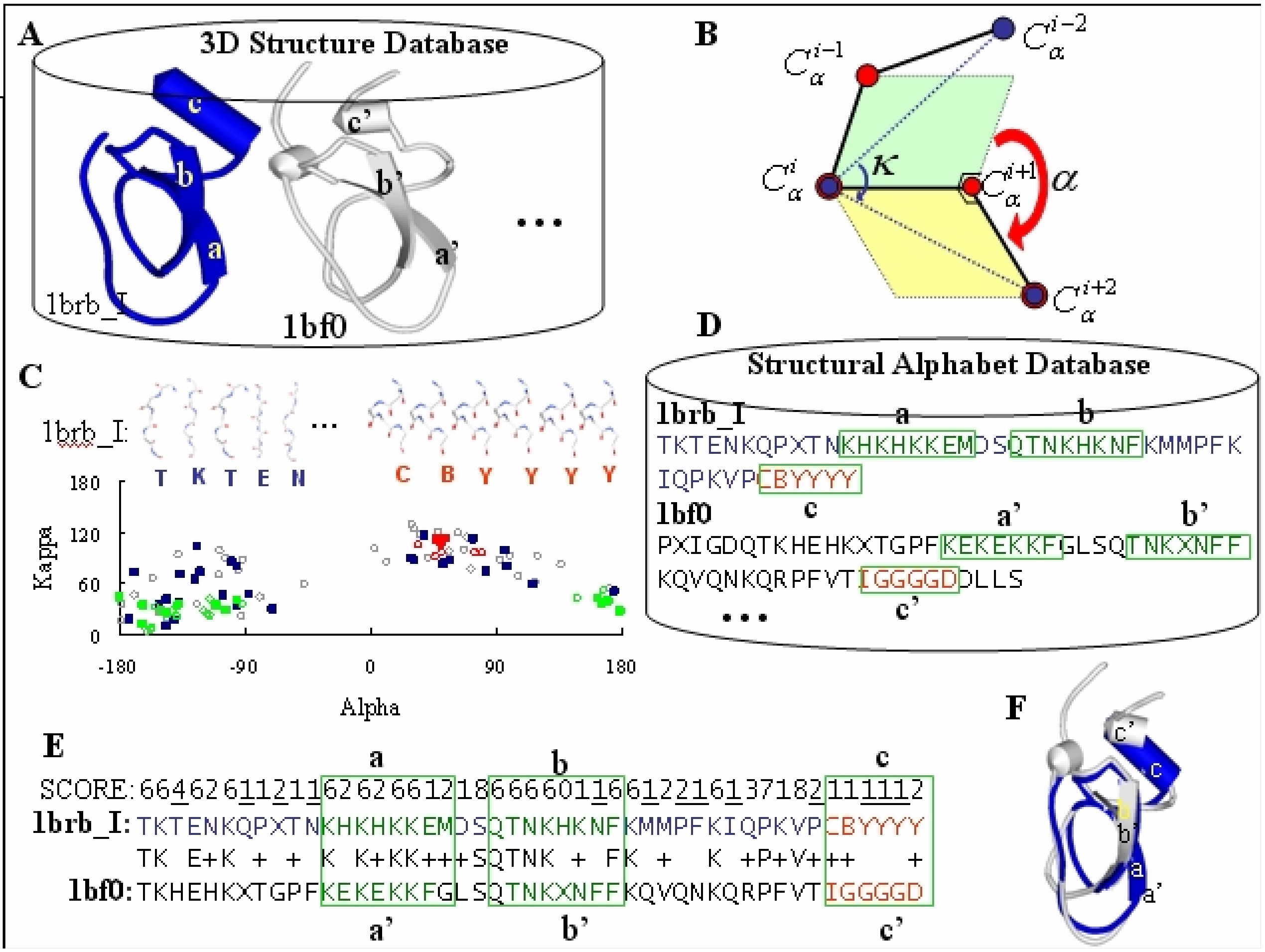
©1999 GARLAND PUBLISHING INC.
A member of the Taylor & Francis Group

SARF: pair of secondary structure

3D-BLAST: unusual definitions

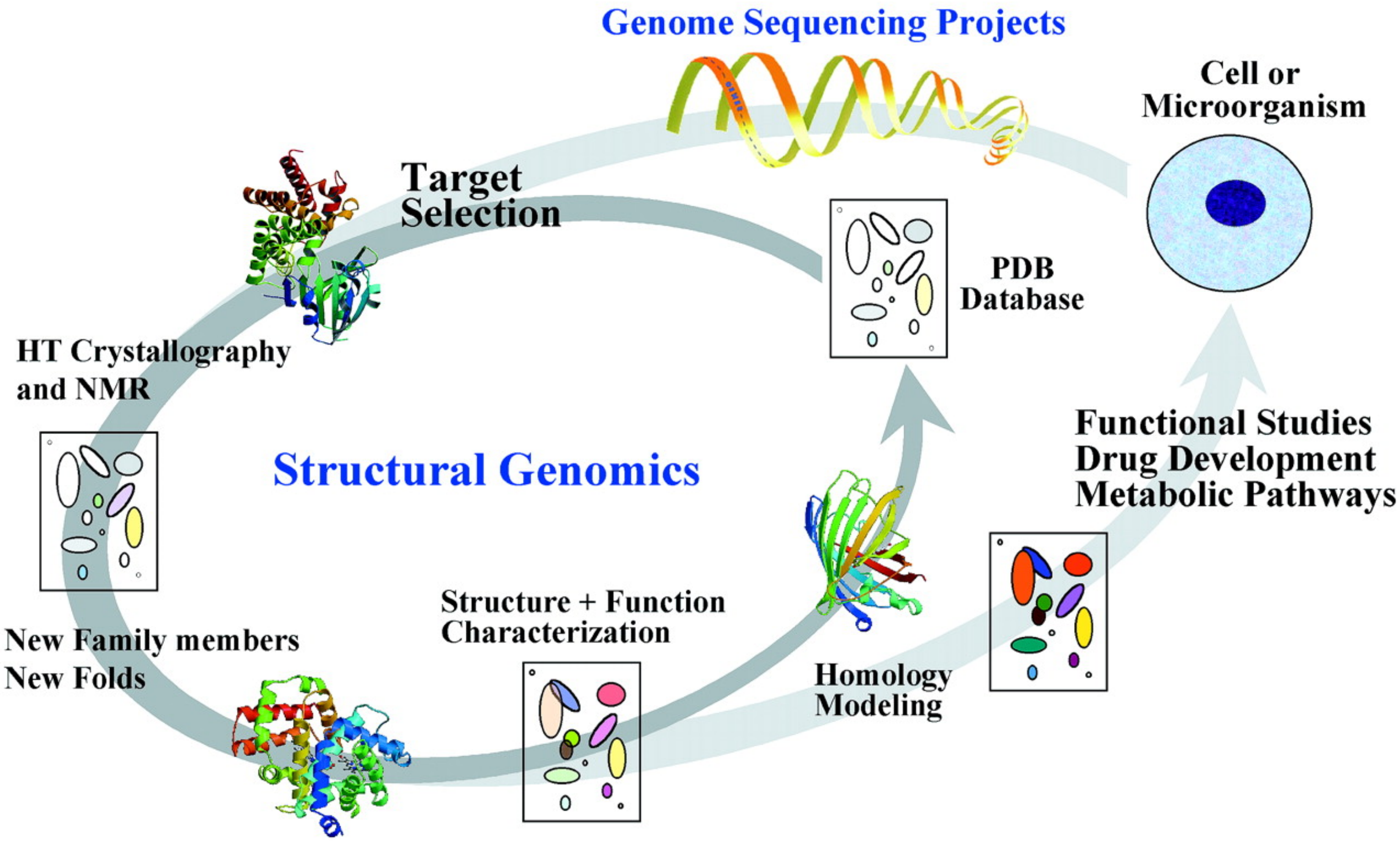
CE: longest path of aligned fragment pairs



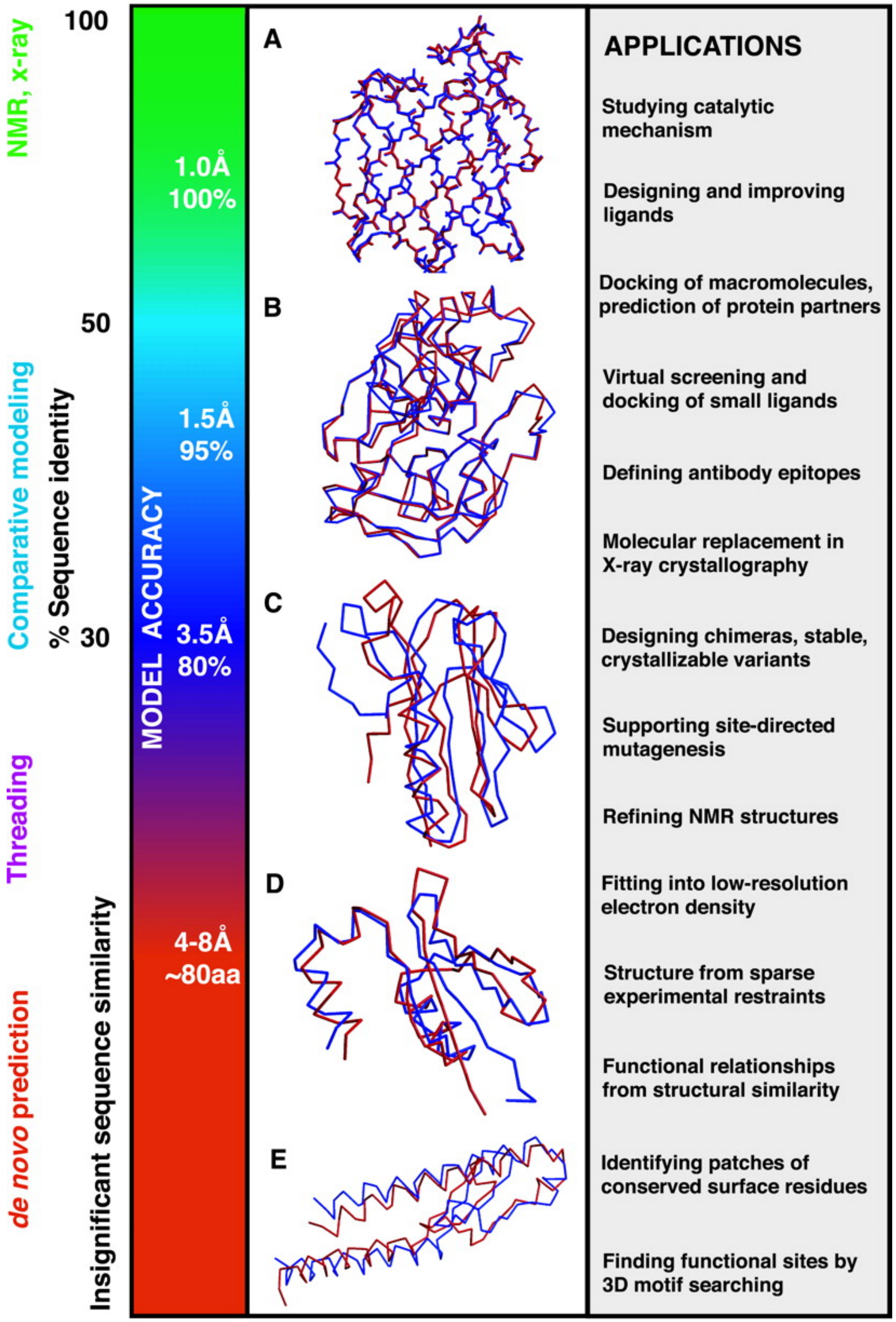

A
3D Structure Database
c
b
a
c'
b'
a'
• • •
1brb_I
1bf0
B
C_α^{i-1}
C_α^{i-2}
C_α^{i}
κ
C_α^{i+1}
α
C_α^{i+2}
C
1brb_I:
T K T E N
…
C B Y Y Y Y
Kappa
180
120
60
0
-180
-90
0
90
180
Alpha
D
Structural Alphabet Database
1brb_I
a
b
TKTENKQPXTNKHKHKKEMDSQTNKHKNFKMMPFK
IQPKVPCBYYYY
1bf0
c
a'
b'
PXIGDQTKHEHKXTGPFKEKEKKFGLSQTNKXNFF
KQVQNKQRPFVTIGGGGDDLLS
c'
• • •
E
a
b
c
SCORE: 6646261121162626612186666011661221613718211112
1brb_I: TKTENKQPXTNKHKHKKEMDSQTNKHKNFKMMPFKIQPKVPCBYYYY
TK E+K + + K K+KK+++SQTNK + FK + K +P+V+++ +
1bf0: TKHEHKXTGPFKEKEKKFGLSQTNKXNFFKQVQNKQRPFVTIGGGGD
a'
b'
c'
F
c'
c
b'
a
a'

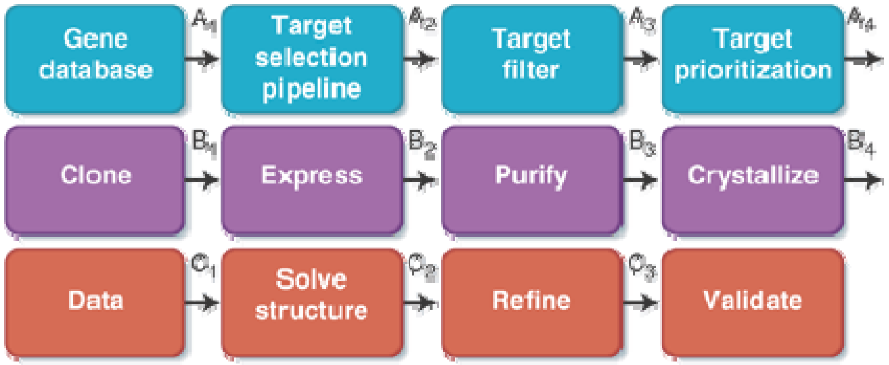
Tertiary structure prediction

- Detailed structural information is essential to model function and to design methods to modulate function
- Experimentally determined structures are used as templates during structural prediction



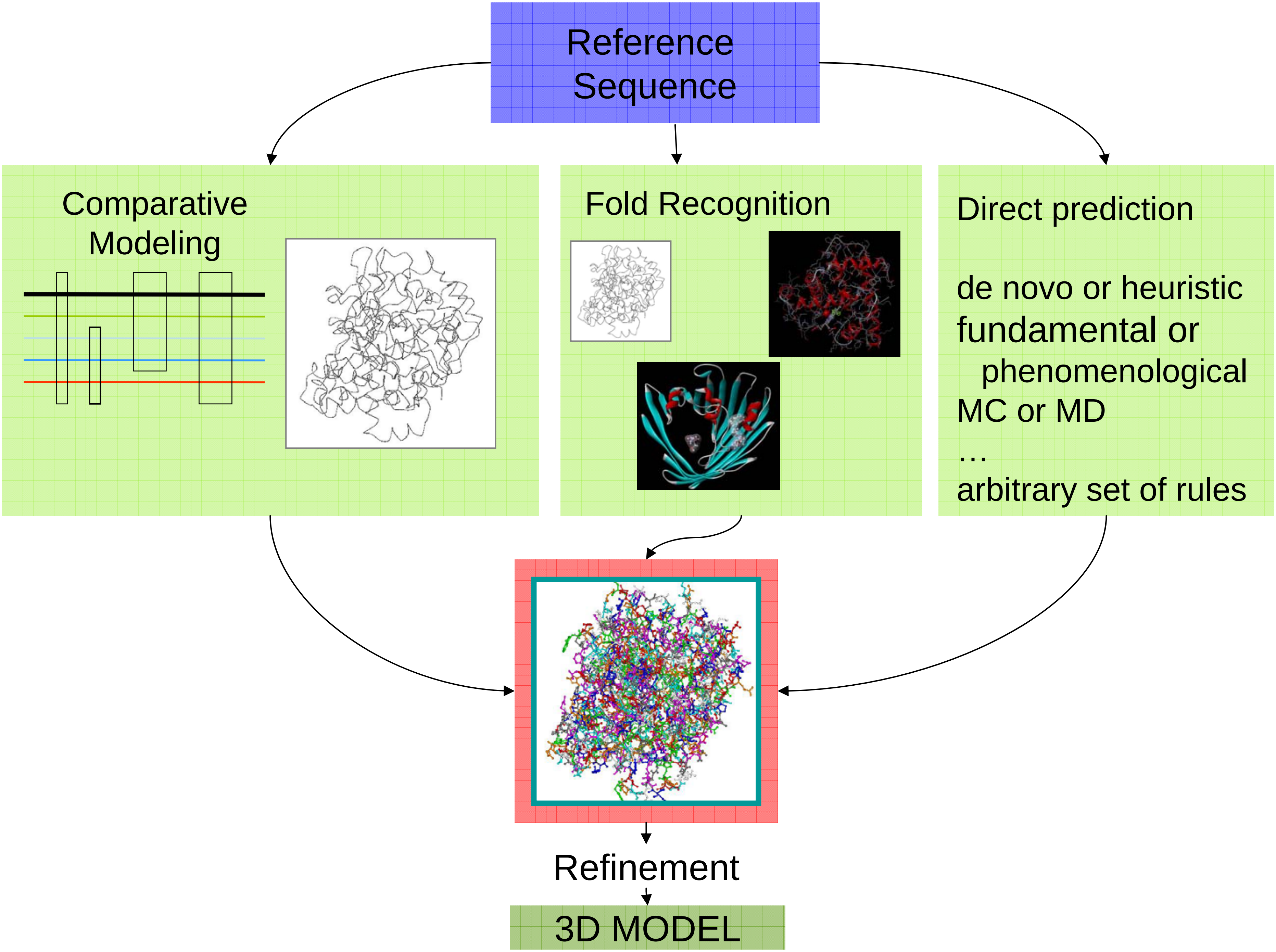
Stevens et al, Science 294, 89-93 (2001)


NMR, x-ray
Comparative modeling
Threading
de novo prediction
100
50
30
% Sequence identity
Insignificant sequence similarity
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
4-8Å
~80aa
A
B
C
D
E
APPLICATIONS
Studying catalytic mechanism
Designing and improving ligands
Docking of macromolecules, prediction of protein partners
Virtual screening and docking of small ligands
Defining antibody epitopes
Molecular replacement in X-ray crystallography
Designing chimeras, stable, crystallizable variants
Supporting site-directed mutagenesis
Refining NMR structures
Fitting into low-resolution electron density
Structure from sparse experimental restraints
Functional relationships from structural similarity
Identifying patches of conserved surface residues
Finding functional sites by 3D motif searching



Stevens & Wilson, Science 291, 519 (2001)

Baker & Sali, Science 294, 93 (2001)



Reference Sequence

Comparative Modeling

Fold Recognition

Direct prediction

de novo or heuristic

fundamental or phenomenological

MC or MD

…

arbitrary set of rules

Refinement

3D MODEL

CASP

Critical Assessment of (Protein) Structure Prediction

Bi-annual competition for testing the current state of structure prediction capabilities

Contestants are given protein sequence and need to submit model structures to be compared against experimental structures

No limit on the technique

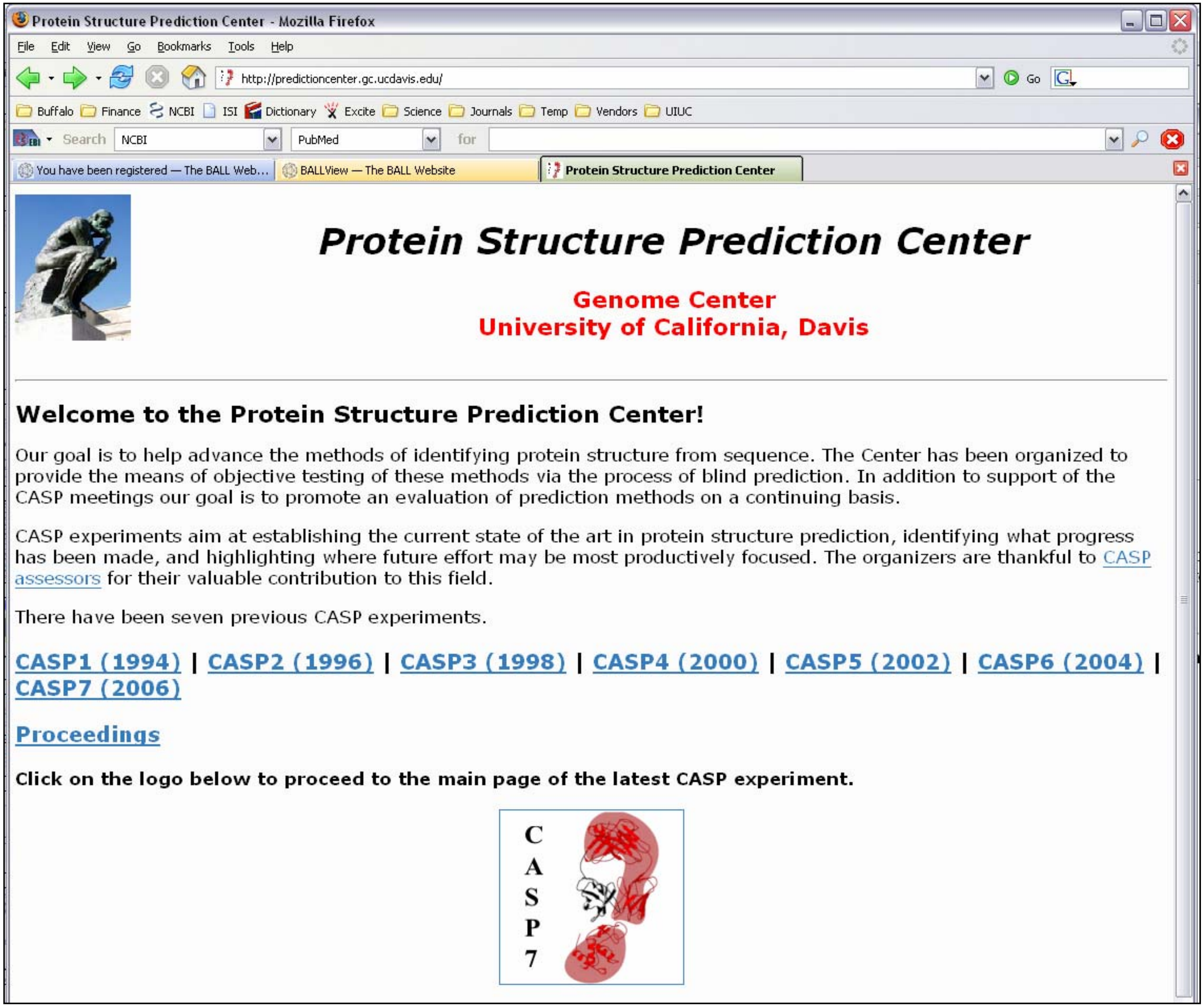
Judging the success of a prediction -- Global and local rmsd

Would like it to be high throughput to cover the full genome

A lot of experimental information cannot be modeled in high throughput, e.g. thermostability and functional site residues

Lack of resolution prevents mutagenesis data, information regarding solvent accessibility (e.g. H/D exchange, fluorescence) to be properly modeled

Domain arrangements (quaternary structure) are also difficult to model


Protein Structure Prediction Center - Mozilla Firefox
File
Edit
View
Go
Bookmarks
Tools
Help
http://predictioncenter.gc.ucdavis.edu/
Go
Buffalo
Finance
NCBI
ISI
Dictionary
Excite
Science
Journals
Temp
Vendors
UIUC
Search
NCBI
PubMed
for
You have been registered — The BALL Web...
BALLView — The BALL Website
Protein Structure Prediction Center
C
A
S
P
7

Protein Structure Prediction Center

Genome Center
University of California, Davis

Welcome to the Protein Structure Prediction Center!

Our goal is to help advance the methods of identifying protein structure from sequence. The Center has been organized to provide the means of objective testing of these methods via the process of blind prediction. In addition to support of the CASP meetings our goal is to promote an evaluation of prediction methods on a continuing basis.

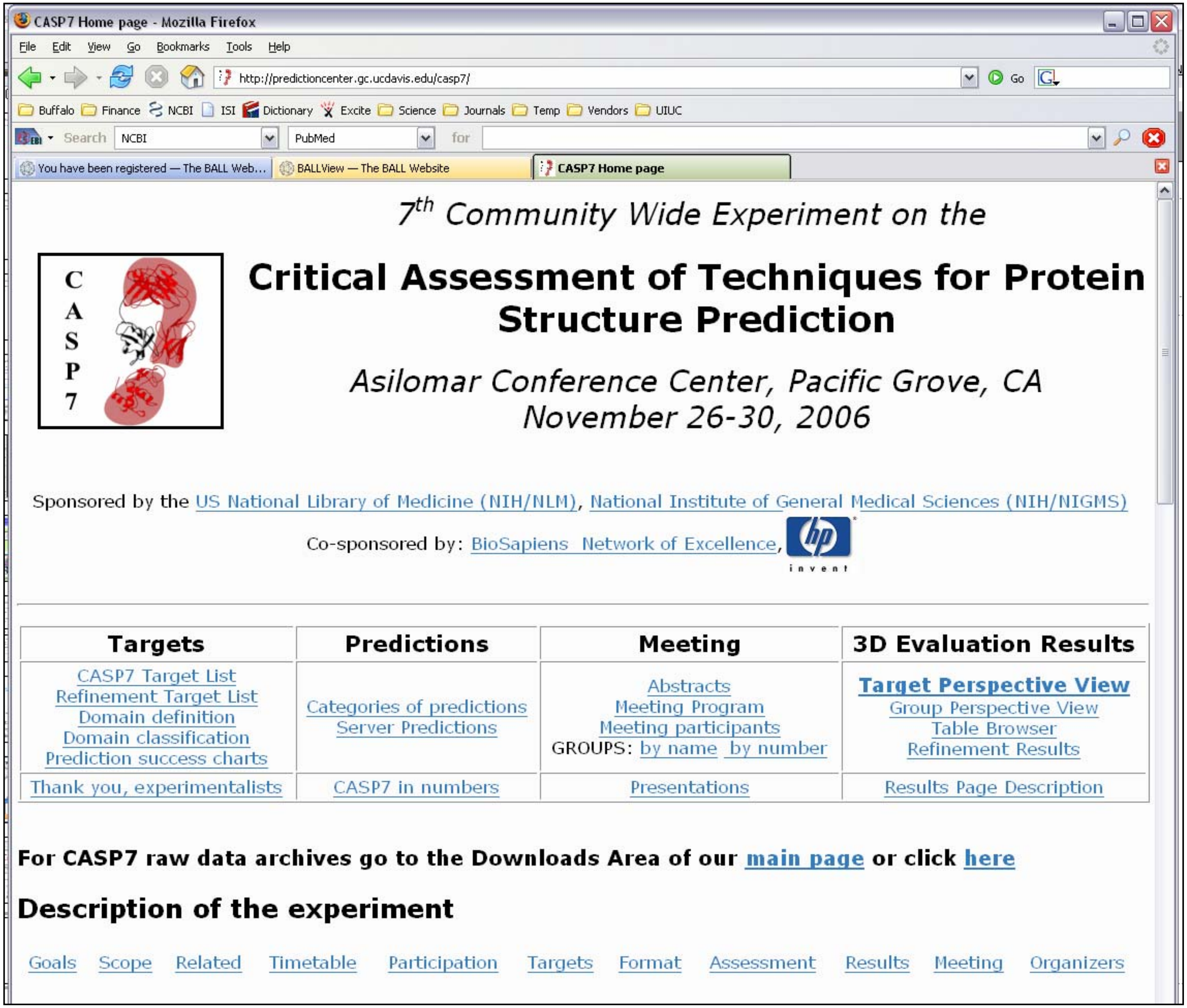
CASP experiments aim at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused. The organizers are thankful to CASP assessors for their valuable contribution to this field.

There have been seven previous CASP experiments.

CASP1 (1994) | CASP2 (1996) | CASP3 (1998) | CASP4 (2000) | CASP5 (2002) | CASP6 (2004) | CASP7 (2006)

Proceedings

Click on the logo below to proceed to the main page of the latest CASP experiment.


CASP7 Home page - Mozilla Firefox
File
Edit
View
Go
Bookmarks
Tools
Help
http://predictioncenter.gc.ucdavis.edu/casp7/
Go
Buffalo
Finance
NCBI
ISI
Dictionary
Excite
Science
Journals
Temp
Vendors
UIUC
Search
NCBI
PubMed
for
You have been registered — The BALL Web...
BALLView — The BALL Website
CASP7 Home page
C
A
S
P
7
invent

7th Community Wide Experiment on the

Critical Assessment of Techniques for Protein Structure Prediction

Asilomar Conference Center, Pacific Grove, CA
November 26-30, 2006

Sponsored by the US National Library of Medicine (NIH/NLM), National Institute of General Medical Sciences (NIH/NIGMS)

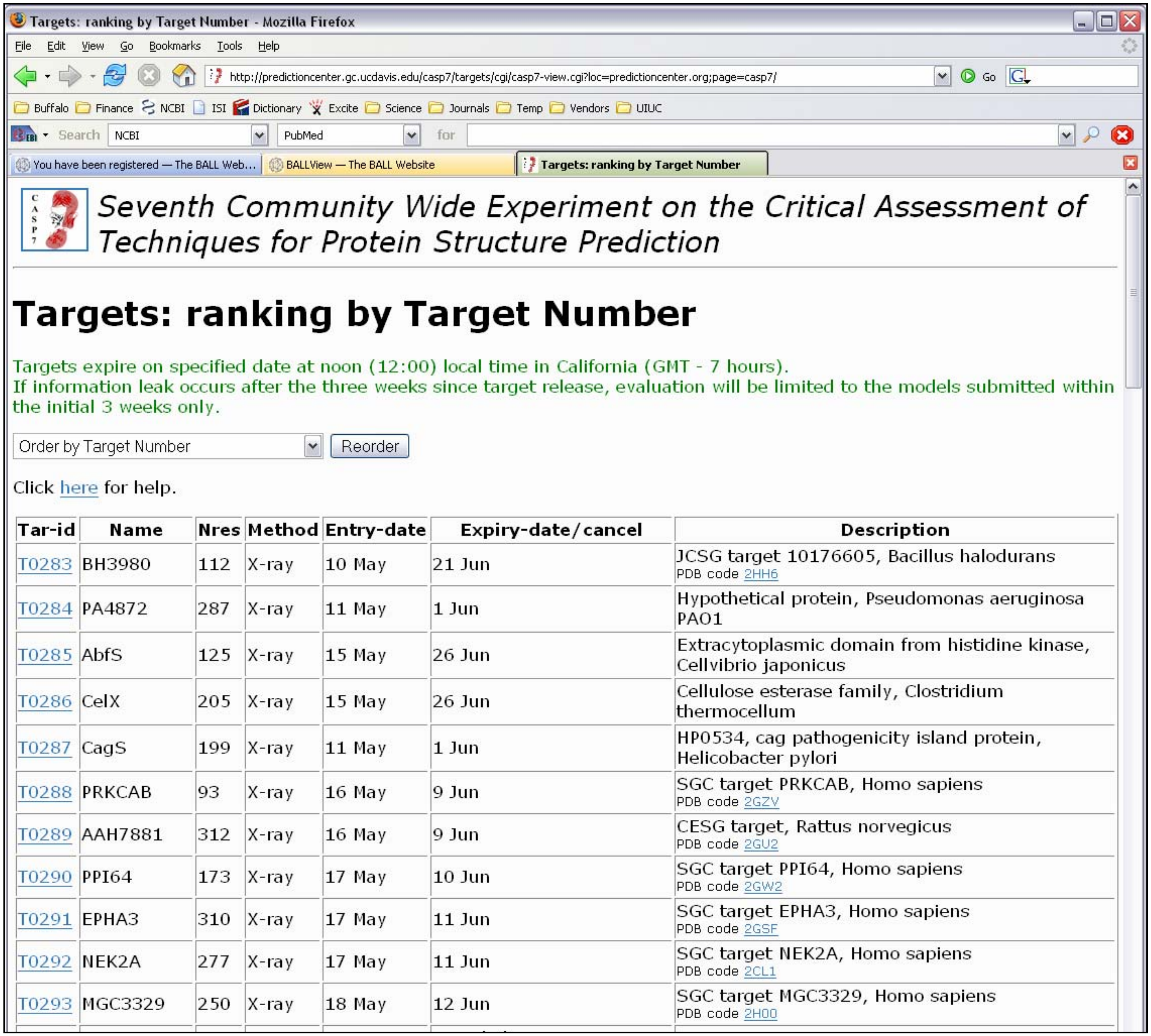
Co-sponsored by: BioSapiens Network of Excellence,

Targets	Predictions	Meeting	3D Evaluation Results
CASP7 Target List Refinement Target List Domain definition Domain classification Prediction success charts	Categories of predictions Server Predictions	Abstracts Meeting Program Meeting participants GROUPS: by name by number	**Target Perspective View** Group Perspective View Table Browser Refinement Results
Thank you, experimentalists	CASP7 in numbers	Presentations	Results Page Description

For CASP7 raw data archives go to the Downloads Area of our main page or click here

Description of the experiment

Goals Scope Related Timetable Participation Targets Format Assessment Results Meeting Organizers


Targets: ranking by Target Number - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://predictioncenter.gc.ucdavis.edu/casp7/targets/cgi/casp7-view.cgi?loc=predictioncenter.org;page=casp7/
Go
Buffalo Finance NCBI ISI Dictionary Excite Science Journals Temp Vendors UIUC
Search NCBI PubMed for
You have been registered — The BALL Web...
BALLView — The BALL Website
Targets: ranking by Target Number
C A S P 7

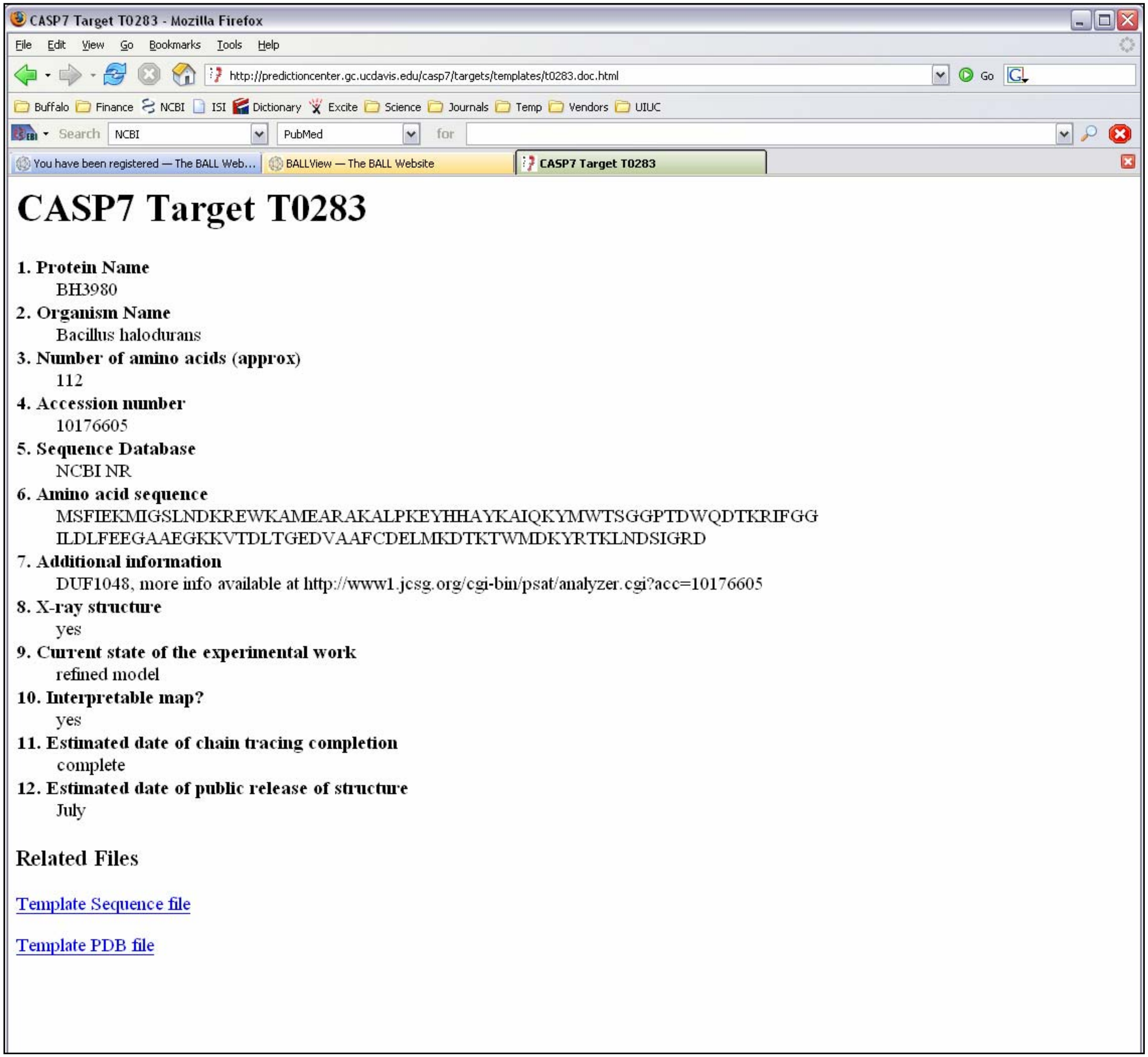
Seventh Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction

Targets: ranking by Target Number

Targets expire on specified date at noon (12:00) local time in California (GMT - 7 hours).
If information leak occurs after the three weeks since target release, evaluation will be limited to the models submitted within the initial 3 weeks only.

Click here for help.

Tar-id	Name	Nres	Method	Entry-date	Expiry-date/cancel	Description
T0283	BH3980	112	X-ray	10 May	21 Jun	JCSG target 10176605, Bacillus halodurans PDB code 2HH6
T0284	PA4872	287	X-ray	11 May	1 Jun	Hypothetical protein, Pseudomonas aeruginosa PAO1
T0285	AbfS	125	X-ray	15 May	26 Jun	Extracytoplasmic domain from histidine kinase, Cellvibrio japonicus
T0286	CelX	205	X-ray	15 May	26 Jun	Cellulose esterase family, Clostridium thermocellum
T0287	CagS	199	X-ray	11 May	1 Jun	HP0534, cag pathogenicity island protein, Helicobacter pylori
T0288	PRKCAB	93	X-ray	16 May	9 Jun	SGC target PRKCAB, Homo sapiens PDB code 2GZV
T0289	AAH7881	312	X-ray	16 May	9 Jun	CESG target, Rattus norvegicus PDB code 2GU2
T0290	PPI64	173	X-ray	17 May	10 Jun	SGC target PPI64, Homo sapiens PDB code 2GW2
T0291	EPHA3	310	X-ray	17 May	11 Jun	SGC target EPHA3, Homo sapiens PDB code 2GSF
T0292	NEK2A	277	X-ray	17 May	11 Jun	SGC target NEK2A, Homo sapiens PDB code 2CL1
T0293	MGC3329	250	X-ray	18 May	12 Jun	SGC target MGC3329, Homo sapiens PDB code 2H00


CASP7 Target T0283 - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://predictioncenter.gc.ucdavis.edu/casp7/targets/templates/t0283.doc.html
Go
Buffalo Finance NCBI ISI Dictionary Excite Science Journals Temp Vendors UIUC
Search NCBI PubMed for
You have been registered — The BALL Web...
BALLView — The BALL Website
CASP7 Target T0283

CASP7 Target T0283

1. Protein Name
BH3980

2. Organism Name
Bacillus halodurans

3. Number of amino acids (approx)
112

4. Accession number
10176605

5. Sequence Database
NCBI NR

6. Amino acid sequence
MSFIEKMIGSLNDKREWKAMEARAKALPKEYHHAYKAIQKYMWTSGGPTDWQDTKRIFGG
ILDLFEEGAAEGKKVTDLTGEDVAAFCDELMKDTKTWMDKYRTKLNDSIGRD

7. Additional information
DUF1048, more info available at http://www1.jcsg.org/cgi-bin/psat/analyzer.cgi?acc=10176605

8. X-ray structure
yes

9. Current state of the experimental work
refined model

10. Interpretable map?
yes

11. Estimated date of chain tracing completion
complete

12. Estimated date of public release of structure
July

Related Files

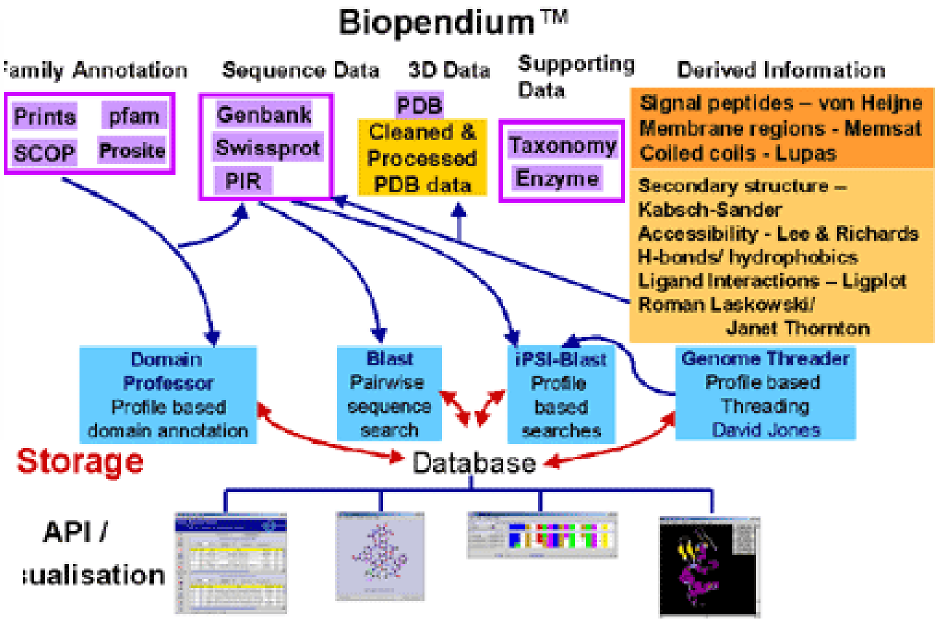
Template Sequence file

Template PDB file

Biopendium™

inpharmatica

http//:www.inpharmatica.com



Biopendium ™

amily Annotation

Sequence Data

3D Data

Supporting Data

Derived Information

Prints

pfam

SCOP

Prosite

Genbank

Swissprot

PIR

PDB

Cleaned & Processed PDB data

Taxonomy

Enzyme

Signal peptides – von Heijne

Membrane regions - Memsat

Coiled coils - Lupas

Secondary structure – Kabsch-Sander

Accessibility - Lee & Richards

H-bonds/ hydrophobics

Ligand Interactions – Ligplot

Roman Laskowski/ Janet Thornton

Domain Professor Profile based domain annotation

Blast Pairwise sequence search

iPSI-Blast Profile based searches

Genome Threader Profile based Threading David Jones

Storage

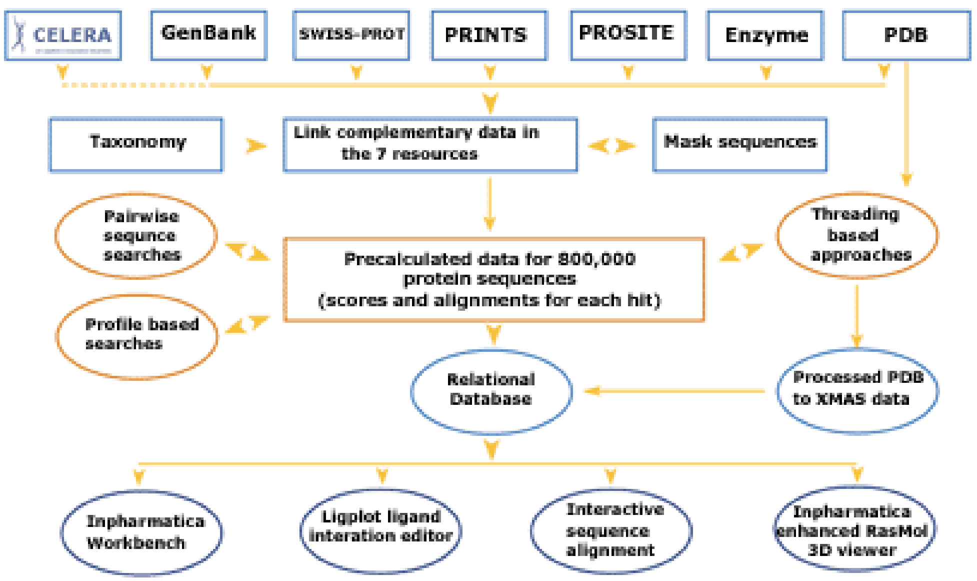
Database

API /

ualisation

The Biopendium™*

Celera Edition Biopendium™ with Celera Human and Mouse Data



CELERA

GenBank

SWISS-PROT

PRINTS

PROSITE

Enzyme

PDB

Taxonomy

Link complementary data in the 7 resources

Mask sequences

Pairwise sequence searches

Profile based searches

Precalculated data for 800,000 protein sequences (scores and alignments for each hit)

Threading based approaches

Relational Database

Processed PDB to XMAS data

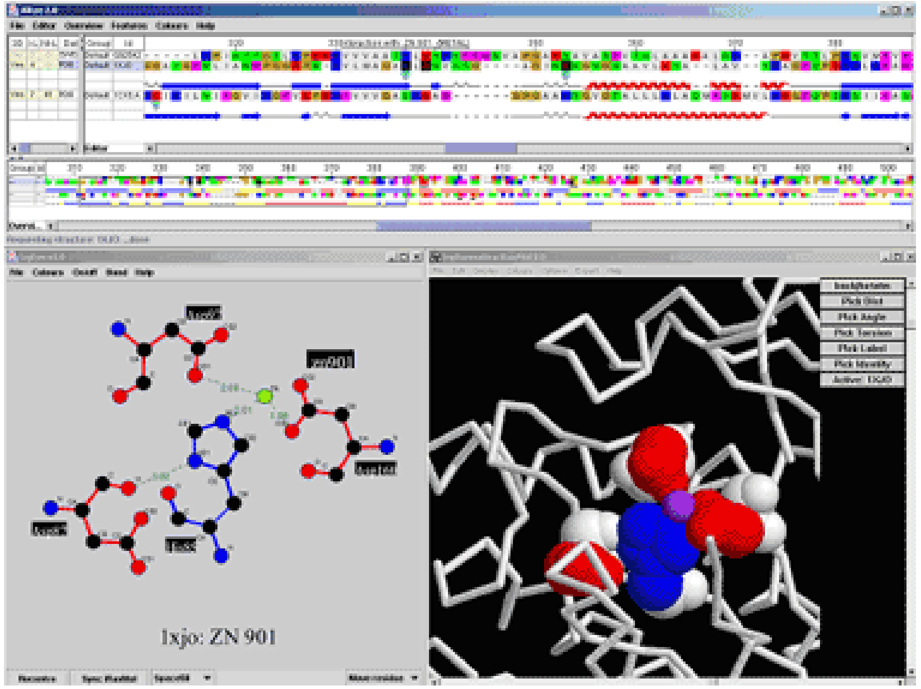
Inpharmatica Workbench

Ligplot ligand interation editor

Interactive sequence alignment

Inpharmatica enhanced RasMol 3D viewer

*Available with or without Celera Data



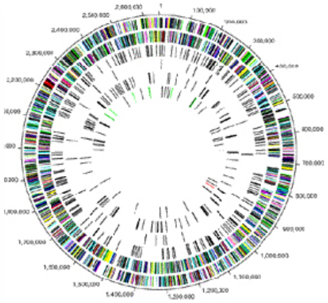
accelrys

Functional Genomics Consortium Datasheet

Accelerating Target Discovery

Genomes Analyzed with GeneAtlas™

Mycoplasma genitalium	*Mycobacterium tuberculosis*
Saccharomyces cerevisiae	*Aquifex aeolicus*
Escherichia coli	*Haemophilus influenzae Rd*
Caenorhabditis elegans	*Methanococcus jannaschii*
Borrelia burgdorferi	*Synechocystis sp.*
Rickettsia prowazekii	*Bacillus subtilis*
Mycoplasma pneumoniae	*Helicobacter pylori*
Archaeoglobus fulgidus	*Pyrococcus horikoshii*
Methanobacterium thermoautotrophicum	*Treponema pallidum*
Chlamydia trachomatis	*Deinococcus radiodurans*
Arabidopsis thaliana (partial)	*Drosophila melanogaster (partial)*
Homo sapiens (partial)	*Vibrio cholerae*



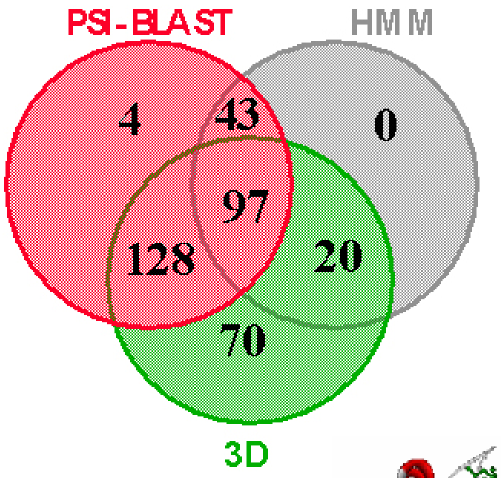
accelrys

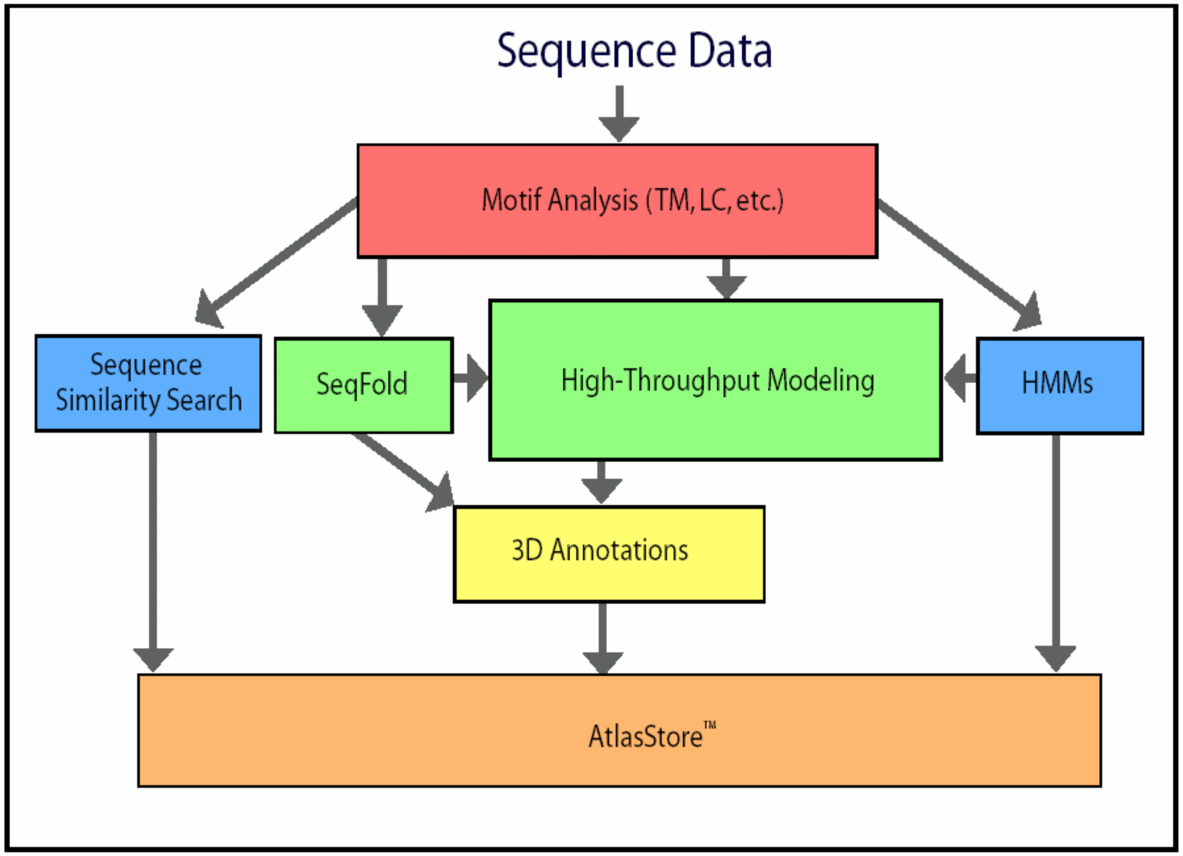
HYSEQ INC.

GEMA Biotech

GENAISSANCE PHARMACEUTICALS

eXegenics





Sequence Data

Motif Analysis (TM, LC, etc.)

Sequence Similarity Search

SeqFold

High-Throughput Modeling

HMMs

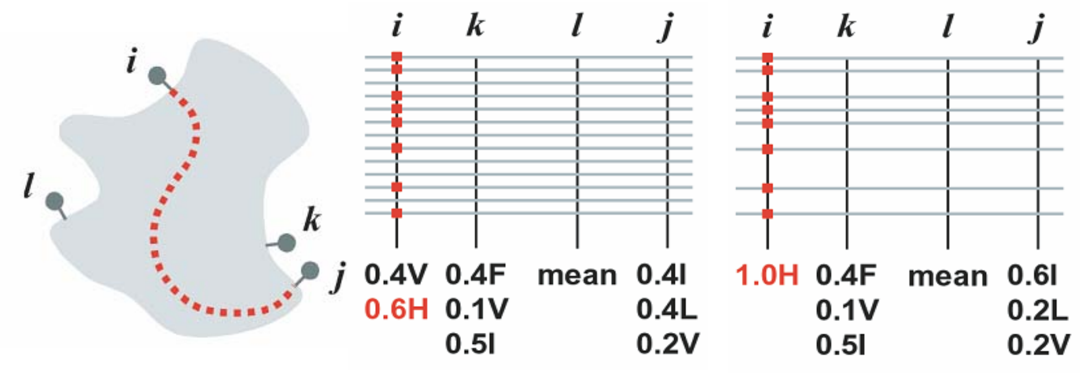
3D Annotations

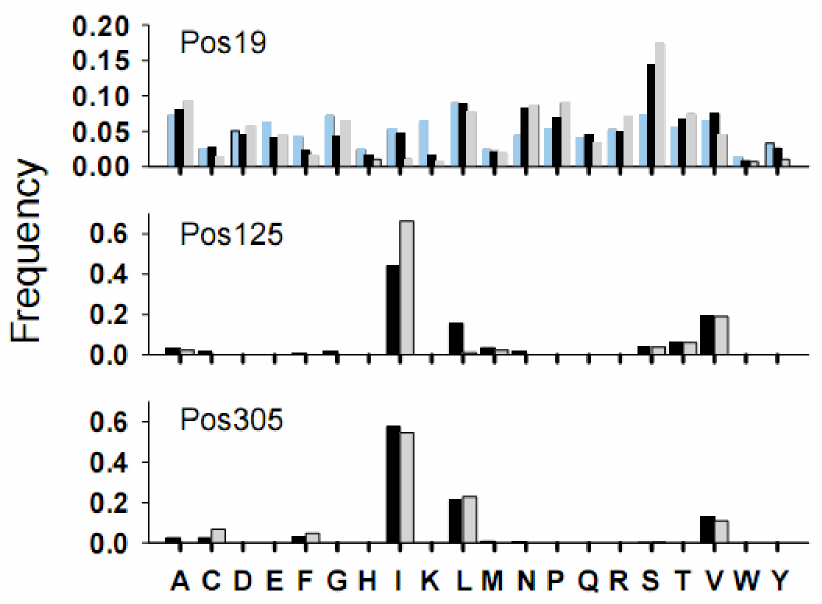
AtlasStore™

Bioinformatics and protein engineering

- Information required for specifying the tertiary structure is contained in the amino acid sequence
- Can we extract the information and use it to specify a protein fold?
- Use statistical information encoded in a multiple sequence alignment

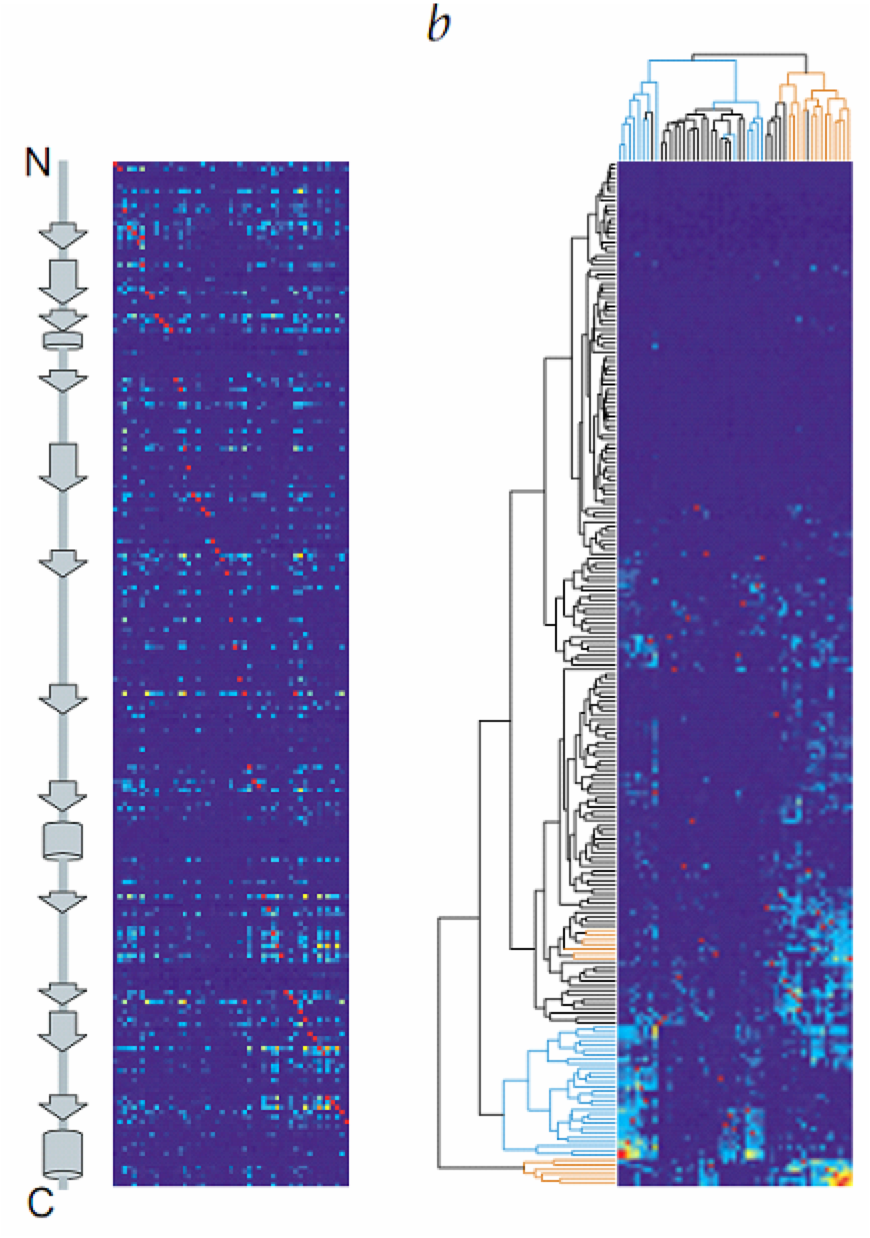
Hypothesis: structural coupled residues would appear more often together than statistically expected

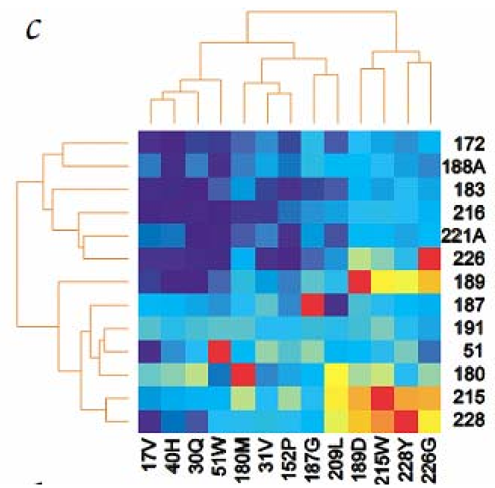


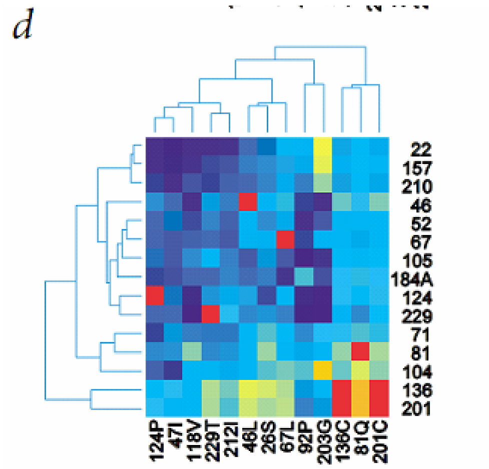


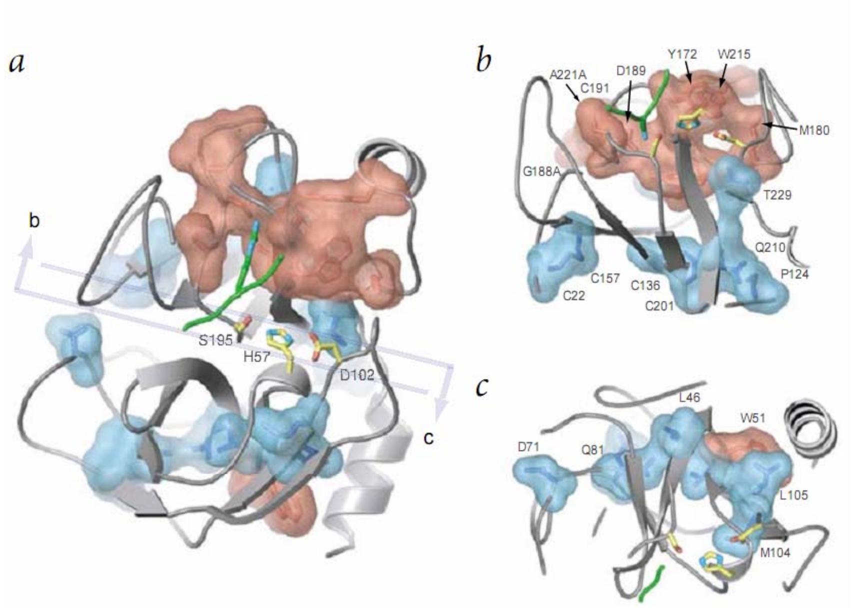
Suel et al, NSB (2003)

Chymotrypsins









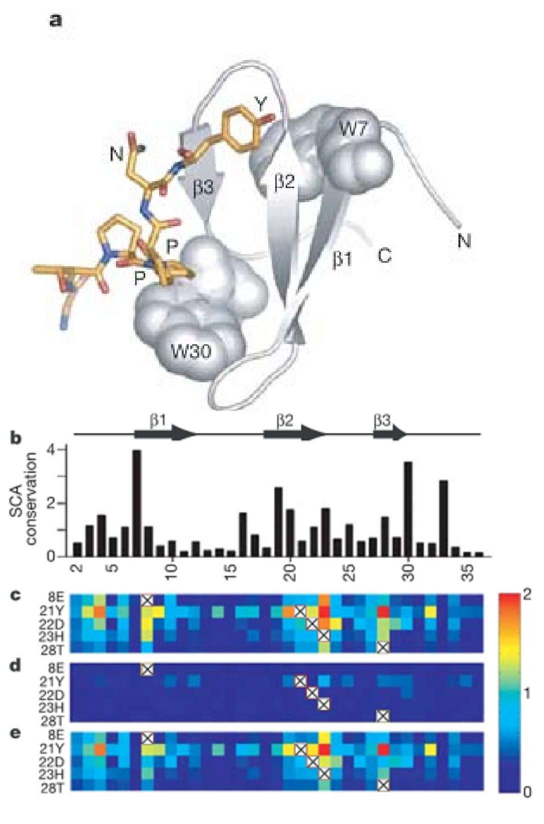
Designing a fold from sequence conservation

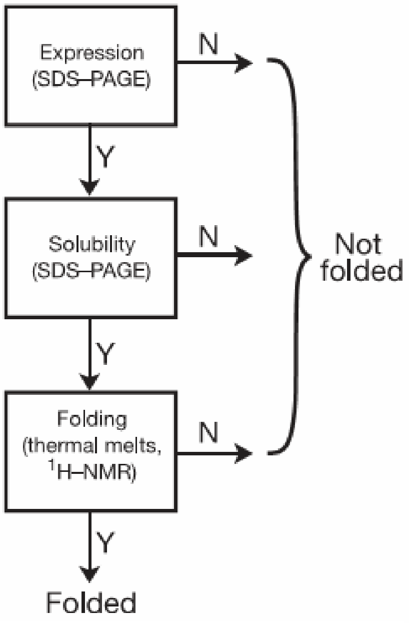
Apply statistical analysis to 120 WW domain proteins to identify which residues are structurally coupled

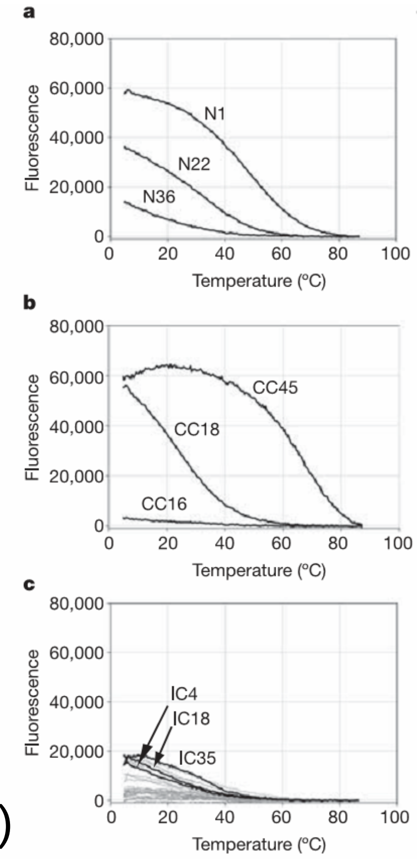
Using simulated annealing Monte Carlo, design sequences that reproduces

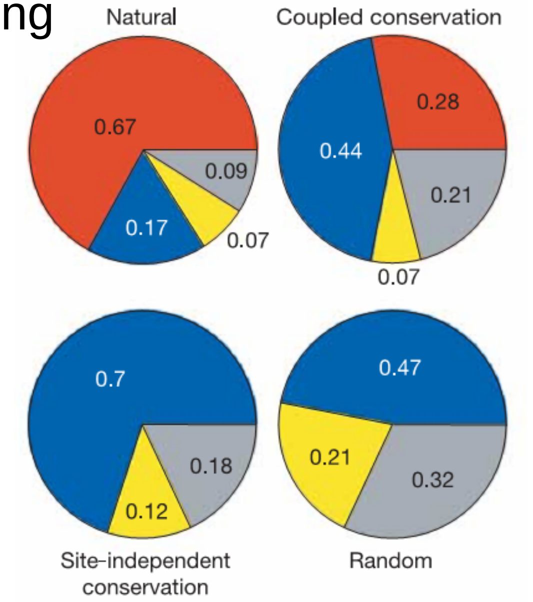
i) intrinsic amino acid distribution at each position, or

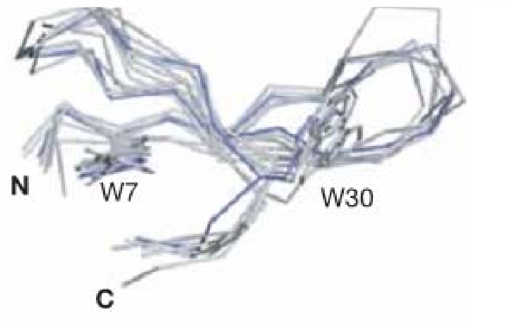
ii) both the sequence conservation and statistical coupling











Socolich et al, Nature 437, 512 (2005)