PredictProtein

online protein sequence analysis and prediction of aspects of protein structure and function

http://ppopen.rostlab.org
PredictProtein.org

Burkhard Rost
Thanks for slides

Guy Yachdav
Laszlo Kajan
Tatyana Goldberg
PP
Background
Overview
Amino acid sequence determines protein 3D structure

Christian Anfinsen
Nobel Prize in Chemistry 1972
Sequence determines structure, determines function

**Sequence**

<table>
<thead>
<tr>
<th>methanobacterium thermoautotrophicum nicotinamide mononucleotide adenyllyltransferase with bound NAD+ (PDB: 1EJ2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEDGYEVPLF</td>
</tr>
</tbody>
</table>

**Function**

"anything that happens to or through a protein"

- chemical
- genetic
- developmental
- cellular
- biochemical
- physiological
Goal: protein sequence annotation

Protein Function

- chemical
- genetic
- developmental
- physiological
- cellular
- biochemical

For us, function is action:
Function = anything that happens to or through a protein (where, what, when)

Protein Structure

GAL4 transcription factor (PDB: 3COQ)
⇒ GAL4: Functional regions highly conserved
⇒ DNA: GAL4 binding site highly conserved

Predict protein structure:
Focus on aspects relevant for function
Growth of protein sequence data

BGI - Beijing Genomics Institute in Shenzhen, China

- <1% of all sequences have an experimental annotation
- <20% of those have a solved 3D structure

Protein annotation and protein-structure gaps widen every day!
PredictProtein (PP) meta-server predicts aspects of...

- Structure
- Function
- Sequence variant effects

6/30/15

Guy Yachdav
23 years of service
OBJECTIVE

serve

computational biologists

and

experimental biologists with little experience in CompBio
Jury returns on structure prediction

Sirs — The jury on structure prediction has returned. The witnesses: the experimental structure of the Src homology 3 (SH3) domain by Musacchio et al., solved by X-ray crystallography, and the accompanying theoretical structure by Benner et al., predicted from the amino-acid sequence. The case: how well did the ETH (see figure) structure-prediction method work in this blind test, performed before the structure became available?

First, the bad news: the tertiary structure prediction was wrong. Benner et al.1 predicted a structure “built from β-strands with a single turn of α-helix lying on one face”. The crystallographers insisted, instead, that “the five strands form two orthogonal β-sheets, as in a β-sandwich”. The only similarity between the predicted and experimental tertiary structures is the presence of β-strands, and that is merely a statement about secondary structure.

Second, the good news: the secondary structure prediction was quite accurate. The protein is correctly predicted to consist mainly of β-strands. Of the five β-strands, four are correctly predicted in about the right sequence position, and one is predicted incorrectly as a helix, a near 80% success rate for segments (ETH in the figure). The success rate is less good when counting how many residues are predicted correctly as helix, strand or loop; only 56%.

Can one do better? For secondary structure, the answer is yes: our novel prediction method2, also tested on the SH3 domain without knowledge of the structure, reached not only 80% for segments, but also 20% for single residues, a good result by current standards. This new method (PHD in the figure) uses information from multiple sequence alignments (as does the ETH method), but has the added advantage of being fully automatic (the ETH method relies in part on human intuition).

We agree with Benner et al. that blind predictions should be made before ex-


Competition of the secondary structure of the SH3 domain determined experimentally and predicted on the basis of multiple sequence alignments by three different methods: A, amino-acid sequence: EXP, crystal structure 1; ETH, de novo prediction of conformation 2; PHD, profile neural network prediction of 250 residues. The only similarity between the predicted and experimental structures is the presence of β-strands, and that is merely a statement about secondary structure.
PredictProtein: online for 23 years

EMBL – 3 tools
(Heidelberg, Germany)

Columbia Univ. – 20 tools
(New York, USA)

TUM >30 tools
(Munich, Germany)

1992  1999  2009
PredictProtein: online for 23 years

EMBL – 3 tools
(Heidelberg, Germany)

Columbia Univ. – 20 tools
(New York, USA)

TUM >30 tools
(Munich, Germany)

1992

1999

2009
PredictProtein: online for 23 years

EMBL – 3 tools
(Heidelberg, Germany)

Columbia Univ. – 20 tools
(New York, USA)

TUM >30 tools
(Munich, Germany)

1992
1999
2009
PredictProtein: online for 23 years

- EMBL – 3 tools
  (Heidelberg, Germany)
- Columbia Univ. – 20 tools
  (New York, USA)
- TUM >30 tools
  (Munich, Germany)
PredictProtein: online for 23 years

EMBL – 3 tools (Heidelberg, Germany)

Columbia Univ. – 20 tools (New York, USA)

TUM >30 tools (Munich, Germany)

1992

1999

2009
PredictProtein: online for 23 years

EMBL – 3 tools
(Heidelberg, Germany)

Columbia Univ. – 20 tools
(New York, USA)

TUM >30 tools
(Munich, Germany)

1992
1999
2009
PredictProtein: online for 23 years

EMBL – 3 tools (Heidelberg, Germany)

Columbia Univ. – 20 tools (New York, USA)

TUM >30 tools (Munich, Germany)
PredictProtein: online for 23 years

- EMBL – 3 tools (Heidelberg, Germany)
- Columbia Univ. – 20 tools (New York, USA)
- TUM >30 tools (Munich, Germany)

Timeline:
- 1992
- 1998
- 1999
- 2009
While other services go away

85% of servers launched in 1995 were down by 2007

10% of servers that get published will be gone next year

PredictProtein server in 1992

VMS machine

- incoming mail received on a VAX computer
- pre-processing of the request:
  - assign unique job identifier
  - extract user network address from mail header and actual sequence to "job_id" file
- send job to queue and confirm request to user
- copy "job_id" file to UNIX file system (rcp) and look or wait for an idle UNIX machine
- give control to UNIX machine by starting a remote procedure call (rpc) on UNIX machine
- wait for job completion

UNIX machines

- check "job_id" file for consistency correct file format?
  - is it a DNA sequence?
  - is it a BINHEX file? .......... 
- run FASTA against the latest release of the SwissProt database 
- extract identifiers of homologous sequences from FASTA run
- run multiple sequence alignment program (MaxHom) against list of SwissProt identifiers and write HSSP output file
- run PHD prediction program on HSSP file
- append output files, copy result to VAX file system and exit

Fig. 3. Procedures performed by the PHD server. The Vax/VMS machine manages the incoming and outgoing mail, and sends the jobs to a cluster of four Unix machines. Here, the CPU intensive processes are executed.
PredictProtein input format in 1992

```
Joe Sequencer, Department of Advanced Protein Research, National University, Timbuktu
joe@fibino.chum.edu
# src homology-3 domain (SH3)
KELVILALYDYQSKSPRETVKKGDLTLTLNSTNKDWWKVEVMDRQ
GFVPAAYVXXLD
```

Fig. 4. Format of file to be sent to “PredictProtein@EMBL-Heidelberg.DE”. Any format different from the one shown will result in an error message being sent back to the reader. The best is necessary to recognize automatically that the sequence will start in the following line. If the server works fine, we do not look at the incoming prediction requests. Thus, messages to PredictProtein will remain unanswered. Instead, address queries or notes to PredictHelp@EMBL-Heidelberg.DE.
1992: simple dependency graph
2015: massively interdependent system

>30 methods

{ 121 packages
  73 libraries
  179 edges
}
Heavy usage of PP
Both web server and packages highly used

1.2 Million visits (~700k visitors)  187 countries

Webserver
510k jobs submitted

Debian Packages
89 packages ran 60 million times

Manuscripts
Cited >800 times
36 patent applications
Accelerate results using algorithms to analyze protein molecules

Advancing research through protein sequence analysis

As a comprehensive resource for protein sequence analysis, PredictProtein incorporates several cutting-edge tools for protein structure and function prediction, including:

- Algorithms providing prediction of:
  - Binding/active sites
  - Sub-cellular localization
  - Domain boundaries
  - Fold recognition
  - Inter-residue contacts
  - Regions lacking regular structure
  - Secondary structure
  - Solvent accessibility
  - Transmembrane, globular and coiled coil regions
  - Disulfide-bonds

Highlights

- Predict the most and least useful gene product targets using computational tools

The problem with traditional proteomics research is that lab experiments are expensive and take too long to produce verified results, even from a limited range of inputs. In the race to find answers, is it possible...
Objective: developers

Intuitive, graphical results for your prediction method: jBio

Laszlo Kajan
Objective: developers & users

'My server should return a figure that people can insert *right into their publications*.'

Guy Yachdav
PP Methods
# PP - Methods overview

<table>
<thead>
<tr>
<th>Type</th>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Database searches</strong></td>
<td><strong>BLAST</strong></td>
<td>Pairwise alignment</td>
</tr>
<tr>
<td></td>
<td><strong>PSI-BLAST</strong></td>
<td>Profile-based alignment</td>
</tr>
<tr>
<td></td>
<td><strong>ps_scan</strong></td>
<td>PROSITE scanning programm</td>
</tr>
<tr>
<td><strong>Prediction of structural features</strong></td>
<td><strong>PROFphd</strong></td>
<td>secondary structure and solvent accesibility</td>
</tr>
<tr>
<td></td>
<td><strong>TMSEG</strong></td>
<td>trans-membrane helices</td>
</tr>
<tr>
<td></td>
<td><strong>COILS</strong></td>
<td>coiled-coil regions</td>
</tr>
<tr>
<td></td>
<td><strong>DISULFID</strong></td>
<td>disulfide bonds</td>
</tr>
<tr>
<td></td>
<td><strong>SEG</strong></td>
<td>low-complexity regions</td>
</tr>
<tr>
<td></td>
<td><strong>NORSnet</strong></td>
<td>disordered regions</td>
</tr>
<tr>
<td><strong>Prediction of functional features</strong></td>
<td><strong>LocTree3</strong></td>
<td>sub-cellular localization</td>
</tr>
<tr>
<td></td>
<td><strong>ConSurf</strong></td>
<td>annotation and visualization of functionally important sites</td>
</tr>
<tr>
<td></td>
<td><strong>SNAP2</strong></td>
<td>effect of SNPs on protein function</td>
</tr>
<tr>
<td></td>
<td><strong>metastudent</strong></td>
<td>Gene Ontology (GO) terms</td>
</tr>
<tr>
<td></td>
<td><strong>ISIS2</strong></td>
<td>protein-protein interaction sites</td>
</tr>
<tr>
<td></td>
<td><strong>SomeNA</strong></td>
<td>Protein-DNA binding sites</td>
</tr>
</tbody>
</table>
PredictProtein in the Cloud

In silico

PredictProtein

experiment

effects of amino acid substitutions

unstructured loops

disulfide bridges

disorder

subcellular localization

localization signals

nuclear localization

region

conformational switches

coiled coils

secondary structure

non-regular secondary structure

accessibility

transmembrane helices

transmembrane boundaries

protein domain

protein-DNA interaction sites

protein-protein interaction sites

protein globularity

homologous sequences

evolutionary profile

protein families

Annotations from prediction methods

Unannotated protein sequences

Guy Yachdav

© Burkhard Rost

rostlab.tum.de

32/70
Secondary structure prediction
PHDsec-PROFsec-ReProf

Solvent accessibility prediction
PROFacc
Notation: protein structure 1D, 2D, 3D
PHD: Neural network & evolutionary information

PROFsec: Evolutionary information + more

B Rost (2001) J Struct Biol 134, 204-18
ReProf

Peter Hoenigschmidt
PHDace

**Local Alignment**

- AAA
- AA
- LLL
- LLL
- ARG
- CYS
- GWV

**Input Local in Sequence**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>L</th>
<th>I</th>
<th>G</th>
<th>S</th>
<th>V</th>
<th>ins</th>
<th>del</th>
<th>cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>33</td>
<td>0</td>
<td>0.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>33</td>
<td>0.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>33</td>
<td>66</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>0</td>
<td>0</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>66</td>
<td>0</td>
<td>0</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>66</td>
<td>0</td>
<td>0.48</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Input Global in Sequence**

- Percentage of each amino acid in protein
- Length of protein
- ΔN-term
- ΔC-term

**Hidden Layer**

- Output Layer
- Input Layer
- First Level Only
**PHDhtm**

**Input local in sequence**

```
A C L I G S V ins del cons
100 0 0 0 0 0 0 0 0 1.17
100 0 0 0 0 0 33 0 0 0.42
0 0 100 0 0 0 0 0 33 0.92
0 0 33 66 0 0 0 0 0 0.74
66 0 0 33 0 0 0 0 0 1.17
0 66 0 0 33 0 0 0 0 0.74
0 0 0 33 0 0 66 0 0 0.48
```

**Input global in sequence**

- Percentage of each amino acid in protein:
  - %AA
  - Length
  - Whole protein

- Protein length:
  - (≤60, ≤120, ≤240, >240)

- Distance:
  - Centre, N-term (≤40, ≤50, ≤20, ≤10)
  - Centre, C-term (≤40, ≤50, ≤20, ≤10)

**Diagram**

- **First level sequence-to-structure**
- **Second level structure-to-structure**
Dynamic programming on NN ‘energy’
Heijne rule: positive inside out
Transmembrane helix: TMSE6G

Predicts:
• Alpha-helical transmembrane proteins
• Position of transmembrane helices
• Membrane topology

Uses Neural Networks

Performance:
Identifies 98% of all transmembrane proteins & 60% membrane topology correctly

Michael Bernhofer

Potassium channel subunit
>2ahyA


Tatyana Goldberg -> over 1000 fewer misclassifications in human
Predict effect of sequence variation: SNAP2

Predicts:
Effect of single amino acid substitution (SNP) on protein function

Uses Neural Networks

Performance:
Two-state accuracy of 83% and an AUC of 0.91

LocTree3 - prediction of localization

- 18 classes
- PSI-BLAST + de novo prediction
- eukaryotes: $Q_{18}>80\%$

Predicted Localization: secreted
Meta-Student GO terms by homology

T Hamp et al & B Rost (2013) BMC Bioinf 14 Suppl 3
Protein-protein interface prediction

Yanay Ofran
Bar Ilan Univ, Israel
Prediction of hot spots for CD4

- alanine scan for V1 domain of CD4 (bound to gp120) 
  (A Ashkenazi et al. & DJ Capon (1990) PNAS 87, 7150)
  red: observed
  purple: predicted
  (Y Ofran & B Rost (2006) PLOS Comp Biol)

- structure:
  observed / predicted

© Burkhard Rost

Protein-DNA/RNA interface prediction: SomeNA

Peter Hönigschmid (2012)
Improvement of DNA- and RNA protein binding prediction.
Diploma Thesis in Bioinformatics, TUM & LMU, Munich
ConSurf

Estimates:
The evolutionary rate in protein families
⇒ constraints on evolutionary drift
⇒ important for structure & function

Uses Bayesian and maximum likelihood methods

Performance:
Accurately maps the patterns of conservation on 3D structures


Interface evolution
Welcome Guy  My Queries  Edit Account  Admin Center  Logout  

<table>
<thead>
<tr>
<th>Description of field (hover over description to get help)</th>
<th>Type the required information into the fields</th>
</tr>
</thead>
<tbody>
<tr>
<td>Your email address*</td>
<td><a href="mailto:gy47@columbia.edu">gy47@columbia.edu</a></td>
</tr>
<tr>
<td>Protein name</td>
<td></td>
</tr>
</tbody>
</table>

### Analysis Methods

<table>
<thead>
<tr>
<th>Analysis Methods</th>
<th>PROFsec</th>
<th>ASP</th>
<th>PROFacc</th>
<th>COILS</th>
<th>PHDhtm</th>
<th>PROFtmb</th>
</tr>
</thead>
<tbody>
<tr>
<td>1D Structure Prediction</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Sequence Motif</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Domain Analysis</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cysteines Disulfide Bonding</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Localization</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disorder</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein Interaction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Threading/Fold Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contact Prediction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

© Burkhard Rost RostLab.

Guy Yachdav TUM
https://www.predictprotein.org
PP supports known facts and suggests new insights

A  ER membrane protein complex 4 from human

B

C  Biological Process Ontology

<table>
<thead>
<tr>
<th>#</th>
<th>GO ID</th>
<th>GO Term</th>
<th>Reliability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GO:0006915</td>
<td>apoptosis</td>
<td>48</td>
</tr>
<tr>
<td>2</td>
<td>GO:0034975</td>
<td>protein folding in endoplasmic reticulum</td>
<td>12</td>
</tr>
</tbody>
</table>

PP technology
Putting More Analysis Power Into PP

Control

• Monitor Quality
• Verify feature set
• Reuse

Ease of Maintenance

• Availability
• Version control
• Bug fixes
Various options for acquiring and running the software

Cloud installation steps

- Obtain cloud account
- Create and **conf** management node
- Create and **conf** worker node(s) from PredictProtein image url
PredictProtein cache (PPCache) 
instantaneous response time

Average run-time for a PredictProtein job -> 39.5 mins

- 25M Pre-calculated PredictProtein entries
- All model organisms
- All UniProt/SwissProt
- ~32% UniProt/TrEMBL

PPCache saves approx. 19k CPU hours a year
PredictProtein cache (PPCache)
instantaneous response time

Average run-time for a PredictProtein job -> 39.5 mins

- 25M Pre-calculated PredictProtein entries
- All model organisms
- All UniProt/SwissProt
- ~32% UniProt/TrEMBL

PPCache saves approx. 19k CPU hours a year
PredictProtein - summary

- Free web server
  - visual results
  - help & tutorial
  - HTML, JSON, XML, and TXT output
- downloadable packages for high-throughput
- PPcache pre-calculated results

Many Thanks to

Guy Yachdav
Laszlo Kajan

RostLab@TUM
PredictProtein - team

The PredictProtein Team

Burkhard Rost, Group Leader
- Founded the PredictProtein service
- Contributed the original methods (PHD, PROF) and first online server
- Continues to support the server by all means possible

Guy Yachdav, Project Leader
- Manages ongoing development and collaborations
- Designed, implemented and maintains the PredictProtein pipeline and online service
- Provides ongoing support

Laszlo Kajan, Director of Software Development
- Lead software development and packaging
- Contributed the Freecontact method
- Created and maintains the PredictProtein Virtual Machine

Tim Karl, System Administration
- Provides ongoing hardware maintenance
- Maintains the server's underlying software and databases

Manfred Roos, Software Developer
- Maintains the PredictProtein Knowledgebase
PredictProtein - team/advise

Scientific Board

Edda Kieppmann, Scientific Board Chair
- Edits Site Content
- Coordinates and provides scientific advice

Maximilian Hecht, PhD Student
- Member of the Scientific Editorial Board
- Contributed the SNAP2 method

Tatyana Goldberg, PhD Student
- Member of the Scientific Editorial Board
- Contributed the LOCtree2 method

Tobias Hamp, PhD Student
- Member of the Scientific Editorial Board
- Contributed the Metastudent method for GO Term prediction

Jonas Reeb, PhD Student
- Member of the Scientific Editorial Board
- Responsible for the transmembranal annotations
PredictProtein - team past

Former Contributors

- Jinfeng Liu, Alumnus
  - Contributed code for the PredictProtein pipeline
  - Contributed the NORS, CHOP & CHOPnet (discontinued) methods

- Yana Bromberg, Alumnus
  - Contributed the Original SNAP method

- Avner Schlessinger, Alumnus
  - Contributed the ProfBval, MD, NorsNet methods

- Yanay Ofran, Alumnus
  - Contributed the ISIS and DISIS methods

- Rajesh Nair, Alumnus
  - Contributed the LocTree method

- Henry Bigelow, Alumnus
  - Contributed the PROFmb method

- Sven Mika, Alumnus
  - Provided the UniqueProt method

- Dariusz Przybylski, Alumnus
  - Contributed the AGAPE method

- Kazimierz Wreszczynski, Alumnus
  - Contributed code and ideas
PredictProtein - team: external & beginning

**External Contributors**

Nir Ben-Tal
- Provides ongoing guidance
- Contributed the ConSurf method

Paolo Fresconi
- Contributed the DISULFIND method

**Original Contributors**

Chris Sander
- Made it possible that PredictProtein went online in 1992, and stayed online until 1998
- Original author of the HSSP method

Reinhard Schneider
- Original author of the HSSP method

Gerrit Vriend
- Helped getting the PredictProtein server online

Antoine de Daruvar
- Helped getting the first PredictProtein server online

Roy Omond
- Hacked some of the original code
Live Demo Presentation
(LOTHAR RICHTER)