Recent Advances in Protein Sequence Analysis

Nick V. Grishin

Howard Hughes Medical Institute, Department of Biochemistry,

University of Texas Southwestern

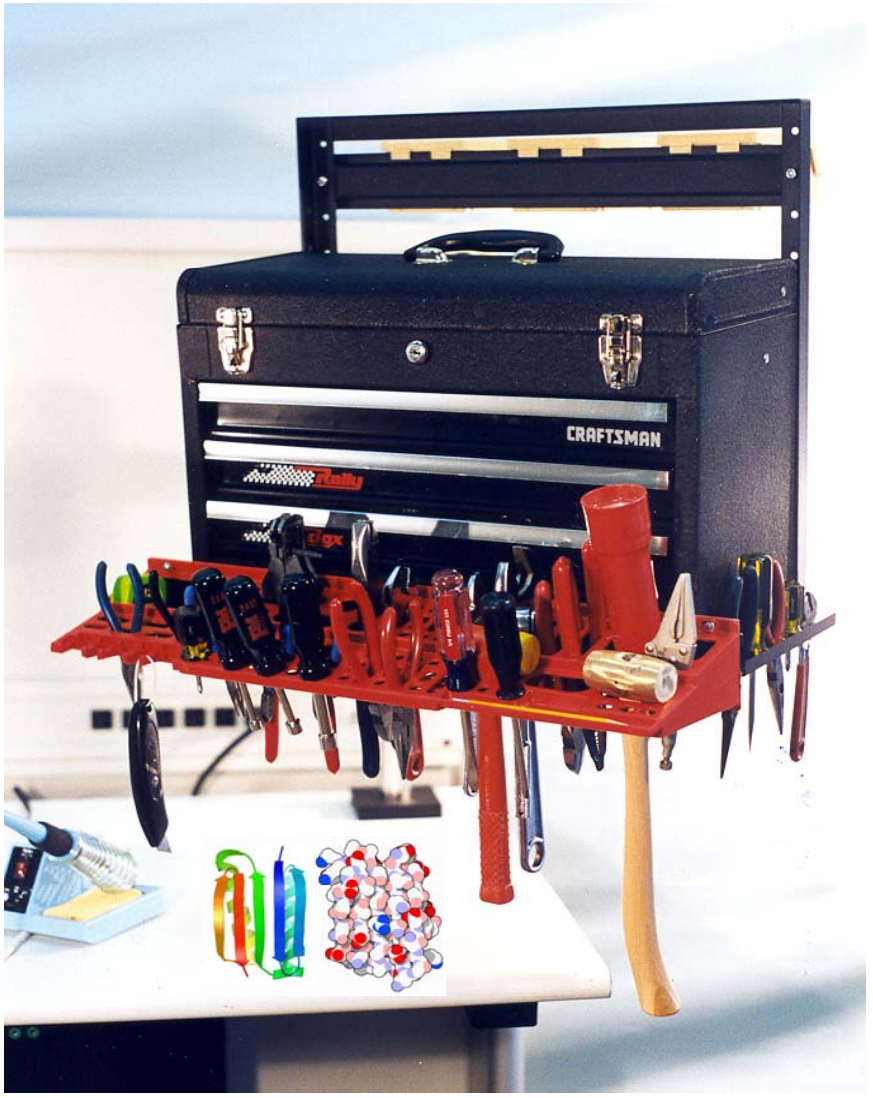
Medical Center at Dallas

Assembling

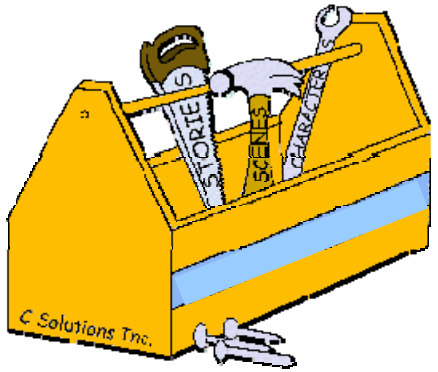
a toolbox

for analysis of

protein molecules



History tour.



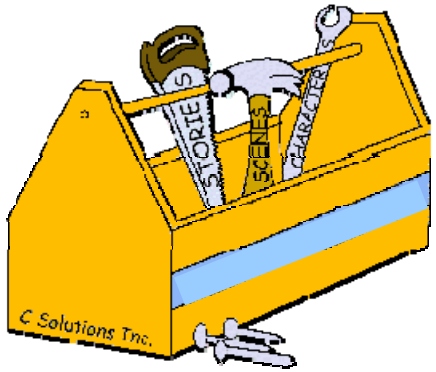
How did it all start?

Question 1: Why is it that educated people (=experts) can understand biological phenomena **so much better** than computers?

Question 2: Why is it that those experts are **so-o-o slow** at what they do best?

Question 3: Why can't these experts **teach** computers to do the job right?

History tour.



How did it all start?

Question 1: Why is it that educated people (=experts) can understand biological phenomena **so much better** than computers?

Maybe they don't

Question 2: Why is it that those experts are **so-o-o slow** at what they do best?

Lazy?

Question 3: Why can't these experts **teach** computers to do the job right?

Snobbish?

We think we are experts.

We are trying to teach computers to give correct answers – and it is hard!

We think we are experts.

We are trying to teach computers to give correct answers – and it is hard!

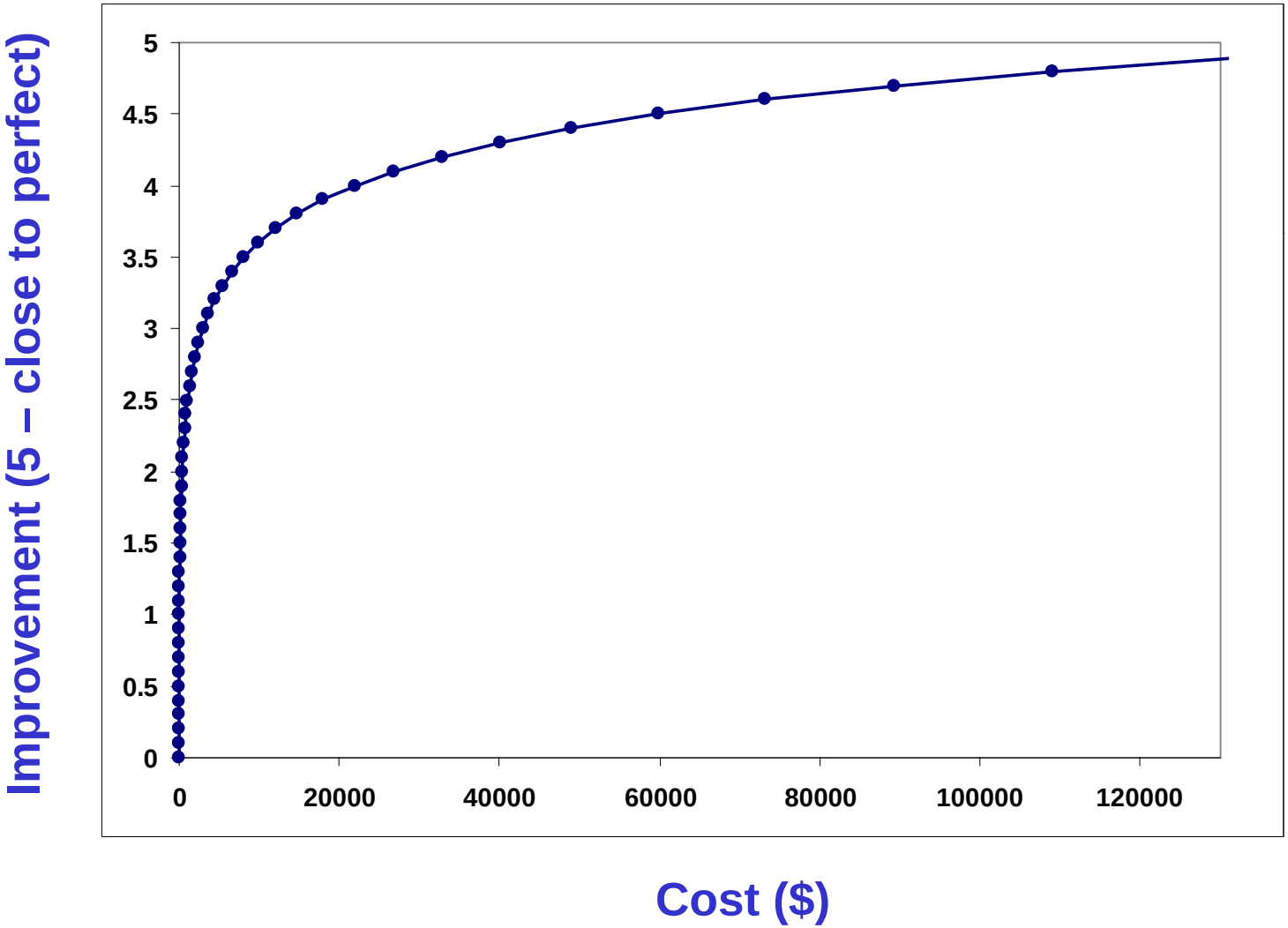
Answers to what questions?

YOU KNOW …

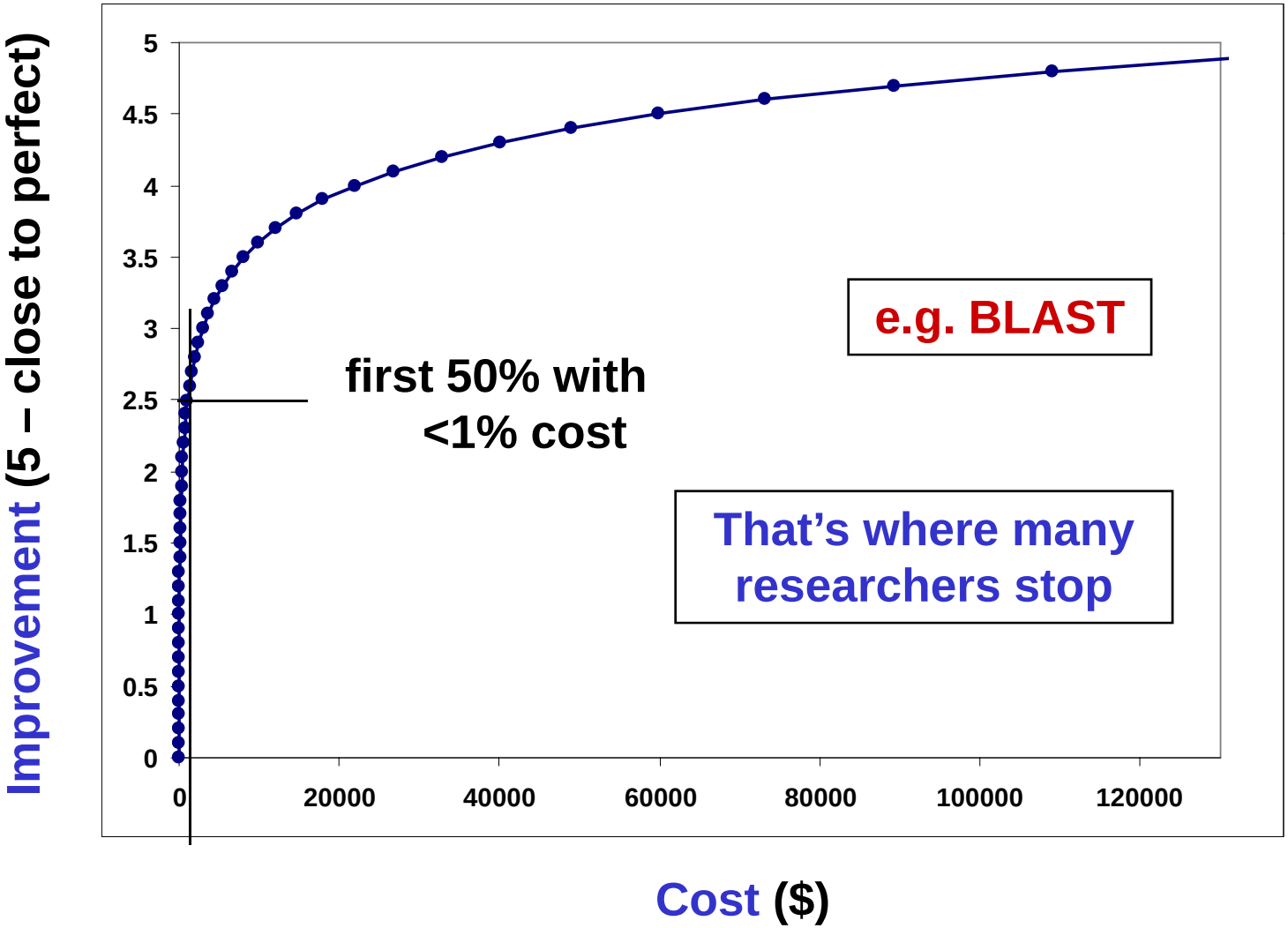
- we have a protein sequence – what is it's 3D structure?
- we have a protein 3D structure – where is the functional site?
- we have 2 sequences – what is their alignment?
- we have many related sequences – what is the tree?

etc. etc. etc.

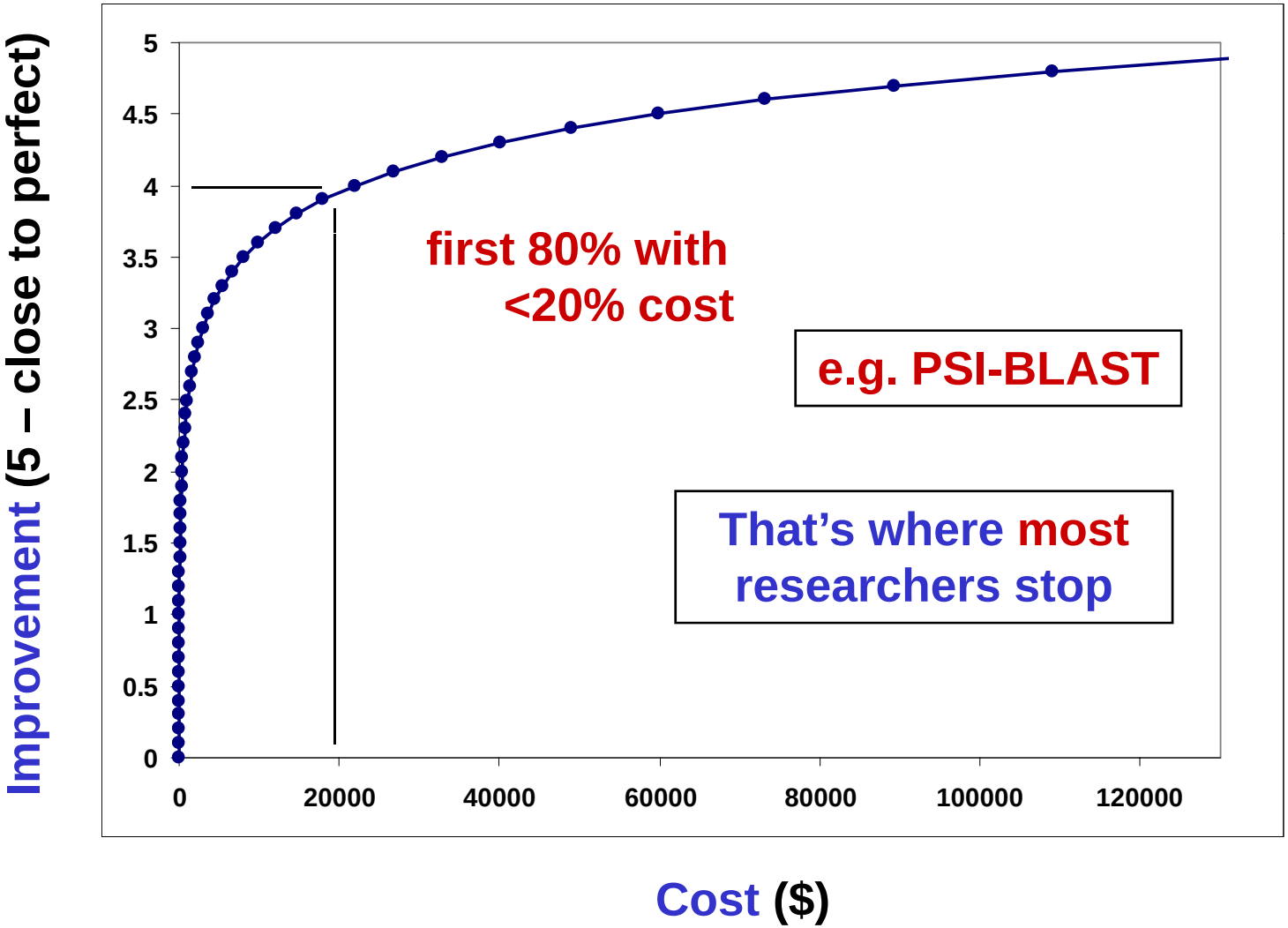
Universal law of science: cost for an increment in improvement increases exponentially with the improvement



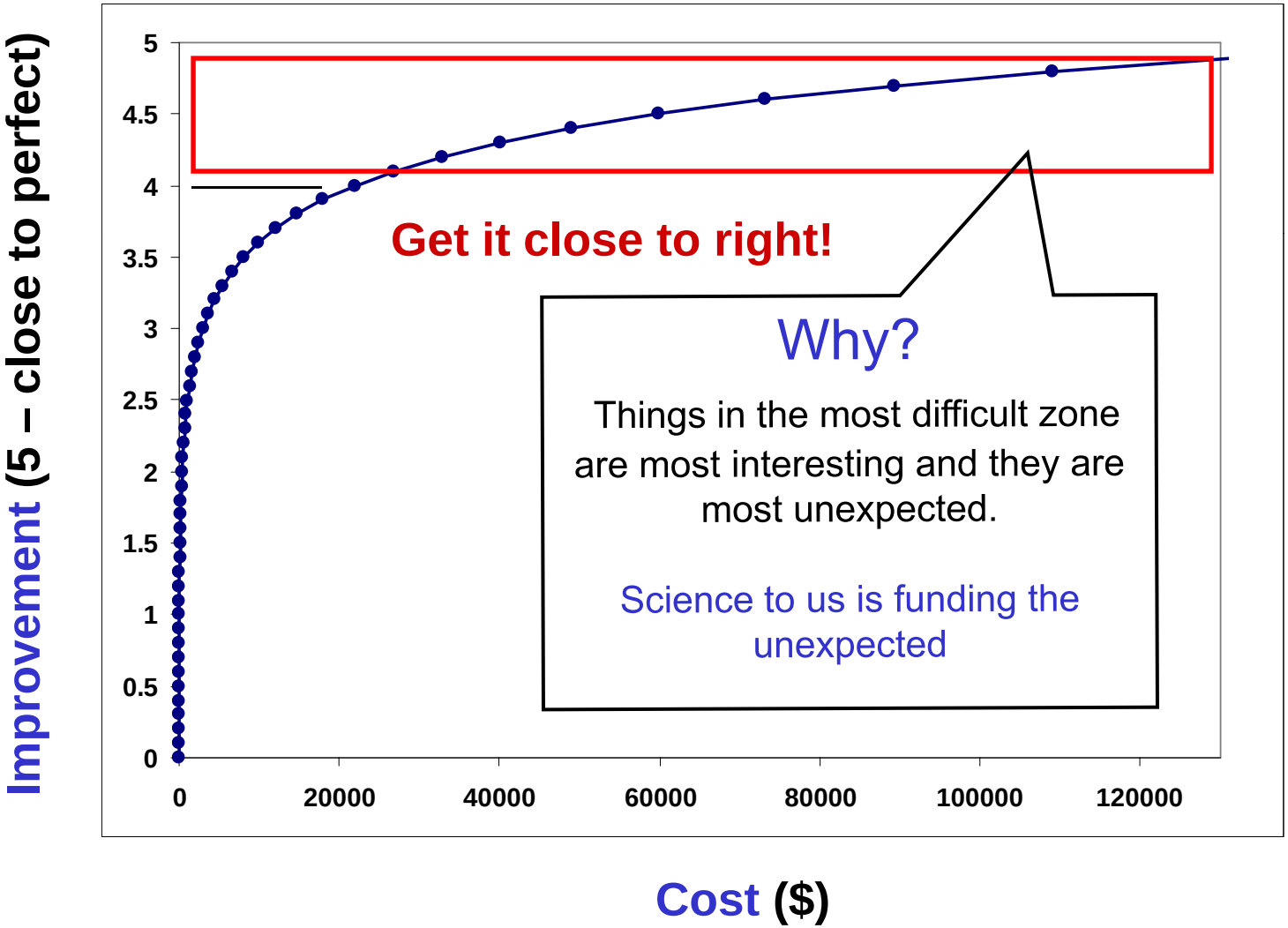
Universal law of science: cost for an increment in improvement increases exponentially with the improvement



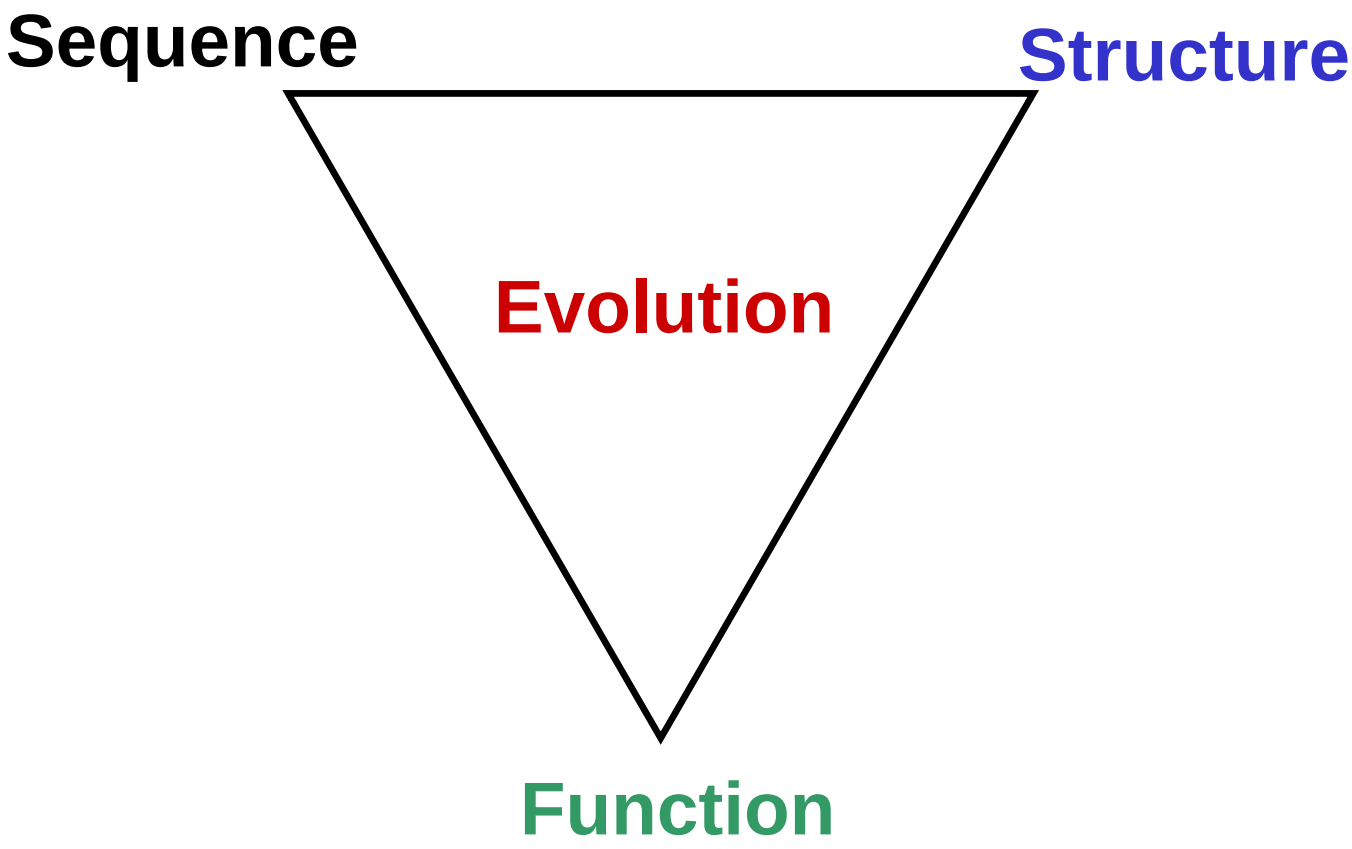
Universal law of science: cost for an increment in improvement increases exponentially with the improvement



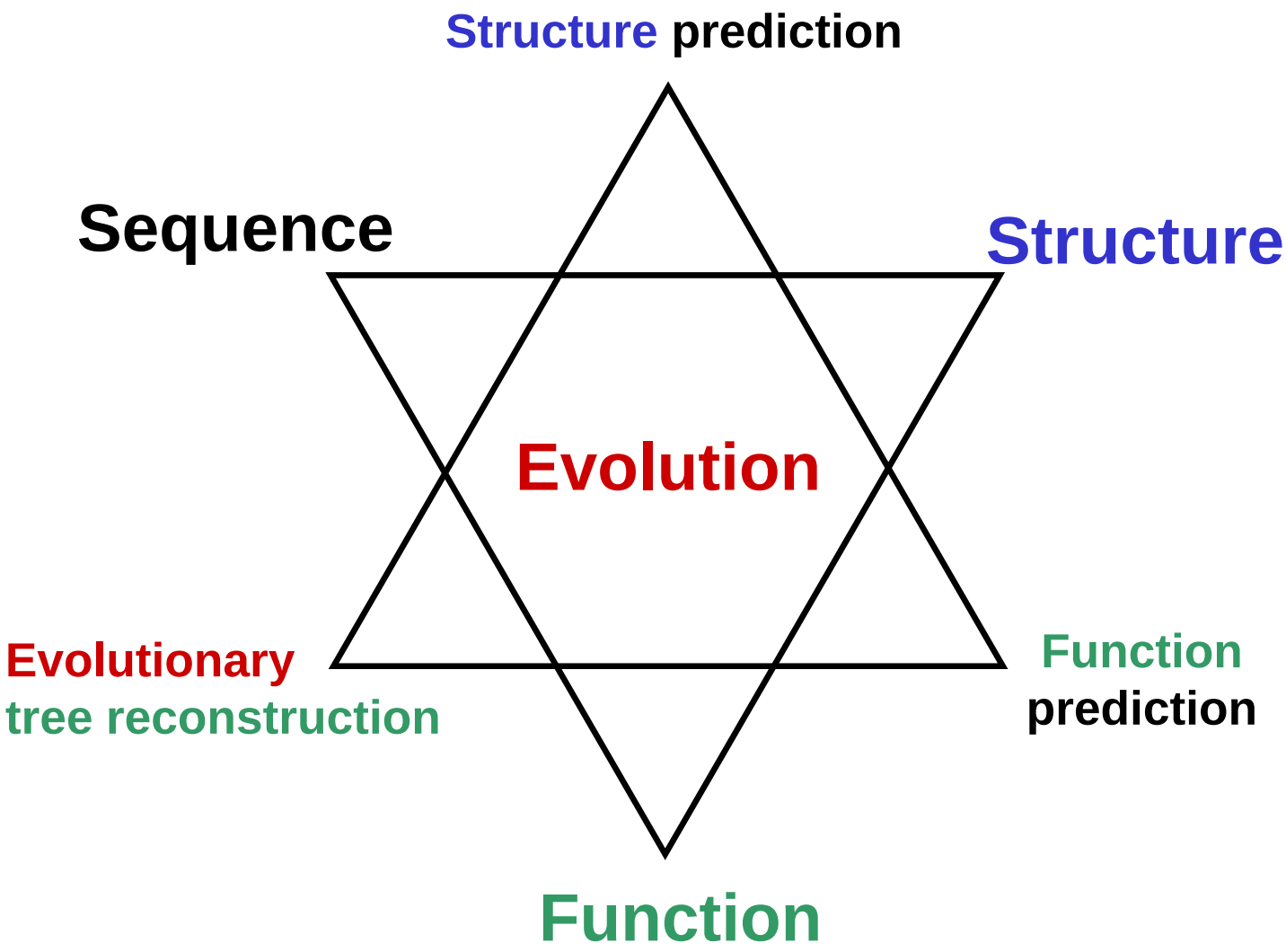
Our Zone


Improvement (5 – close to perfect)
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
20000
40000
60000
80000
100000
120000
Get it close to right!
Why?
Things in the most difficult zone are most interesting and they are most unexpected.
Science to us is funding the unexpected
Cost ($)

Why do we need many tools?



Why do we need many tools?



Main tools in the toolbox

Sequence analysis tools:

- Alignment of alignments and alignment similarity search;
  - plain sequence alignments;
  - alignments with predicted sec.str.
- Multiple sequence alignment;
- Sequence space visualization.

Structure analysis tools:

- Secondary structure delineation;
- Pattern-matching structure similarity search;
- Structure alignment.

Function prediction tools:

- Prediction of functional sites
  - universally important sites;
  - functional specificity sites.
- Evolutionary tree and ancestral sequence reconstruction.

Today's agenda

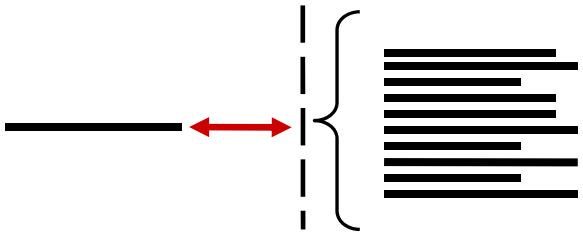
1. COMPASS: Search for similarity between families
2. PROMALS: Multiple sequence alignment

1. COMPASS:
Search for similarity between families

Ruslan Sadreyev

Comparison of multiple alignments improves similarity detection

Sequence-sequence (e.g. BLAST)

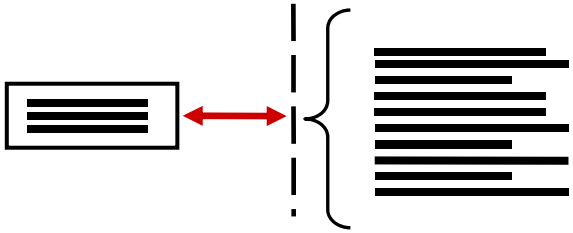


QGVEGPKPAIKLRA

vs.

RVAGMKPRFVRSVKIVHR

Alignment-sequence (e.g. PSI-BLAST)

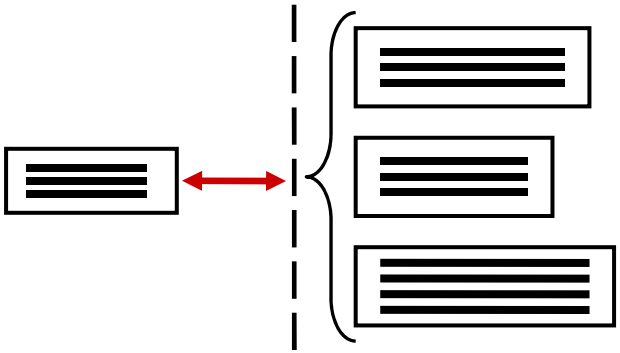


QGVEGPKPAIKLRA
EGLEGPASRFRVTV
KKVDGPPV-SRMTT

vs.

RVAGMKPRFVRSVKIVHR

Alignment-alignment (e.g. COMPASS)



QGVEGPKPAIKLRA
EGLEGPASRFRVTV
KKVDGPPV-SRMTT

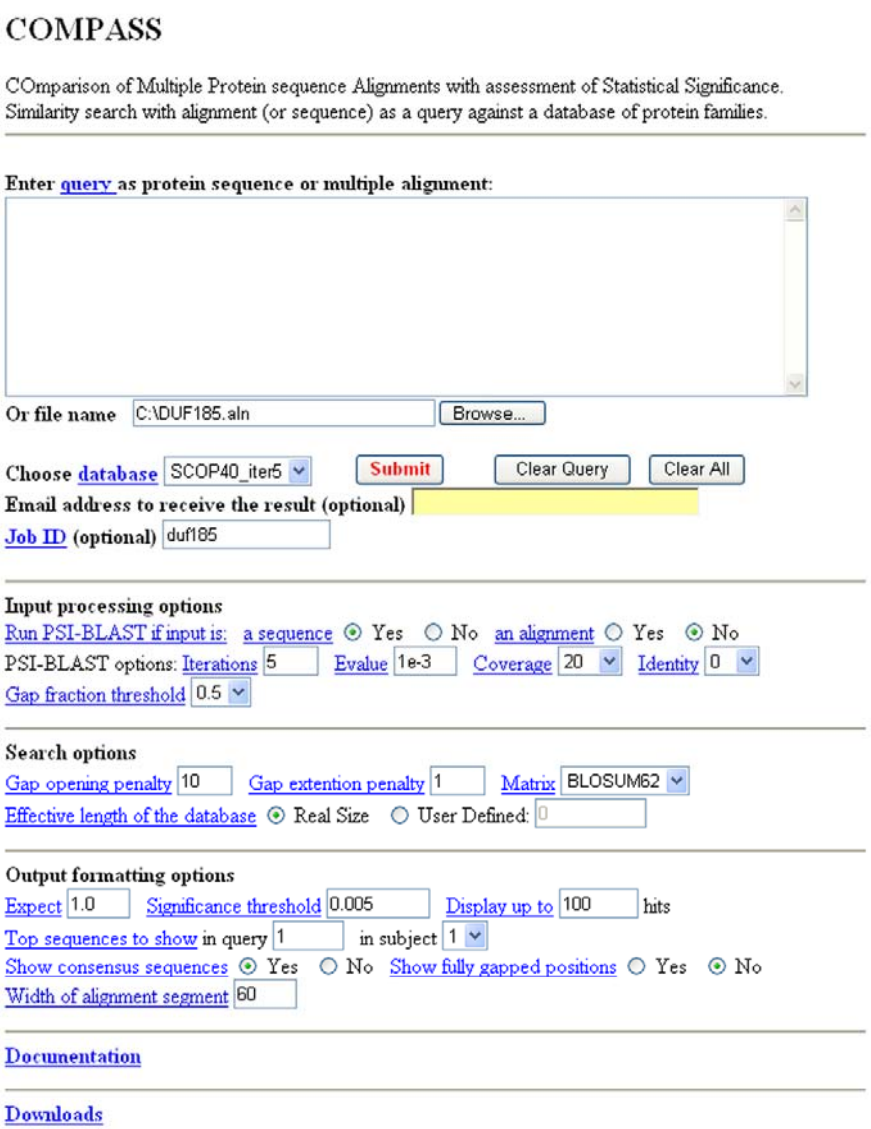
vs.

RVAGMKPRFVRSVKIVHR
IIRASKPKFTRSVTI-HR
QLVGSKPKFTRTLVT-HR

COMPASS web server

http://prodata.swmed.edu/compass

COMPASS:
a method for
COmparison of
Multiple
Protein
Alignments with
assessment of
Statistical
Significance



Sadreyev and Grishin (2003) *JMB*, 326: 317

Recent changes: 2007

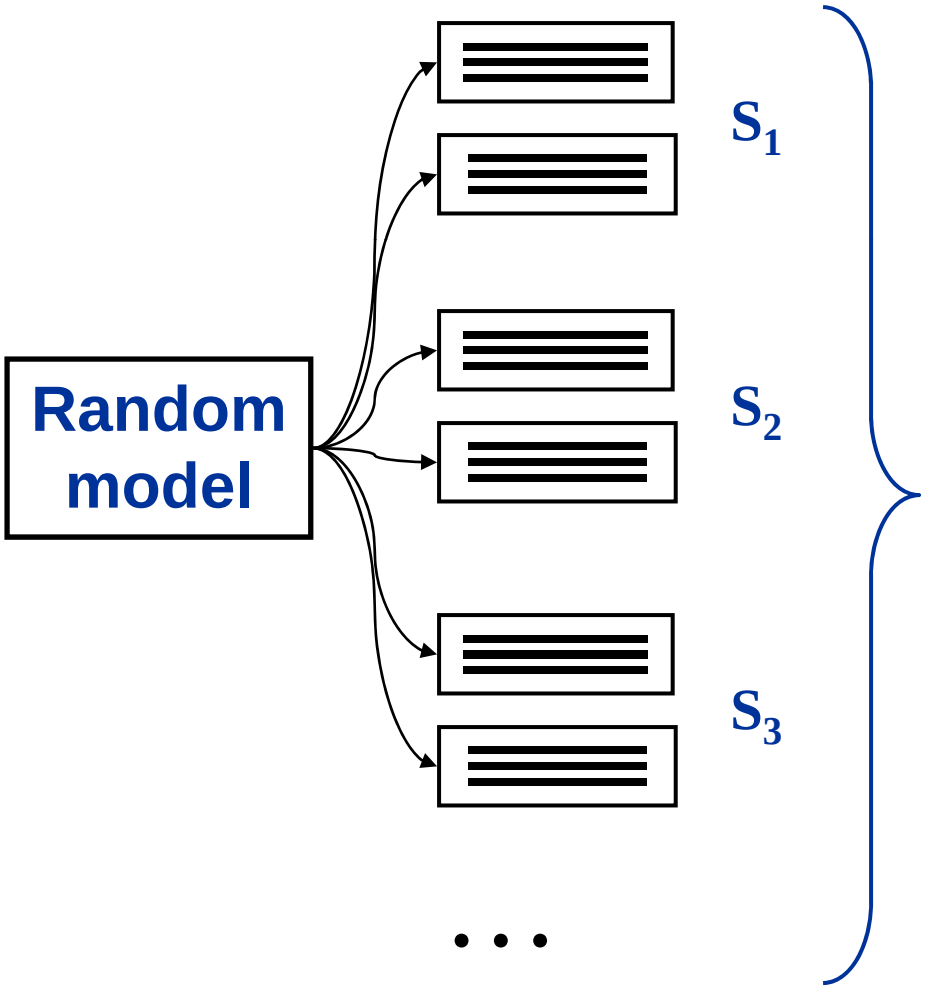
1. New random model for profiles

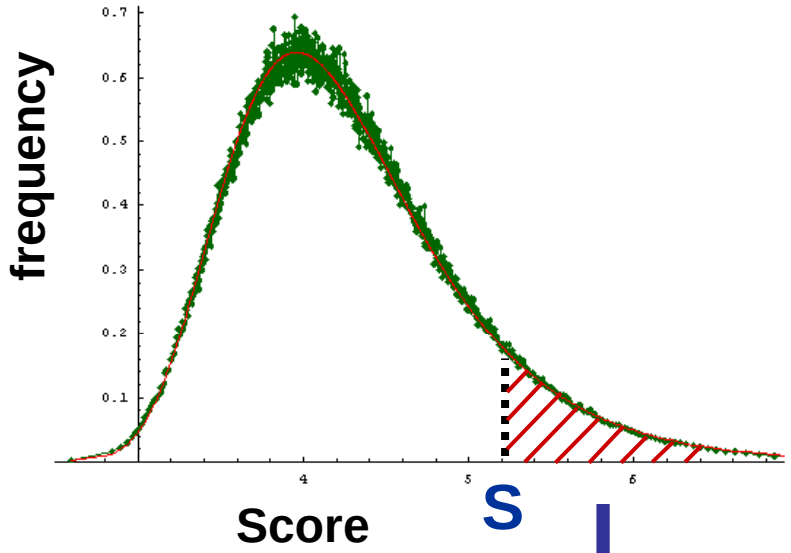
2. New distribution to describe scores

Estimates of statistical significance are based on a random model of alignment comparison

Random decoy profiles

Score distribution





E-value

Old random model

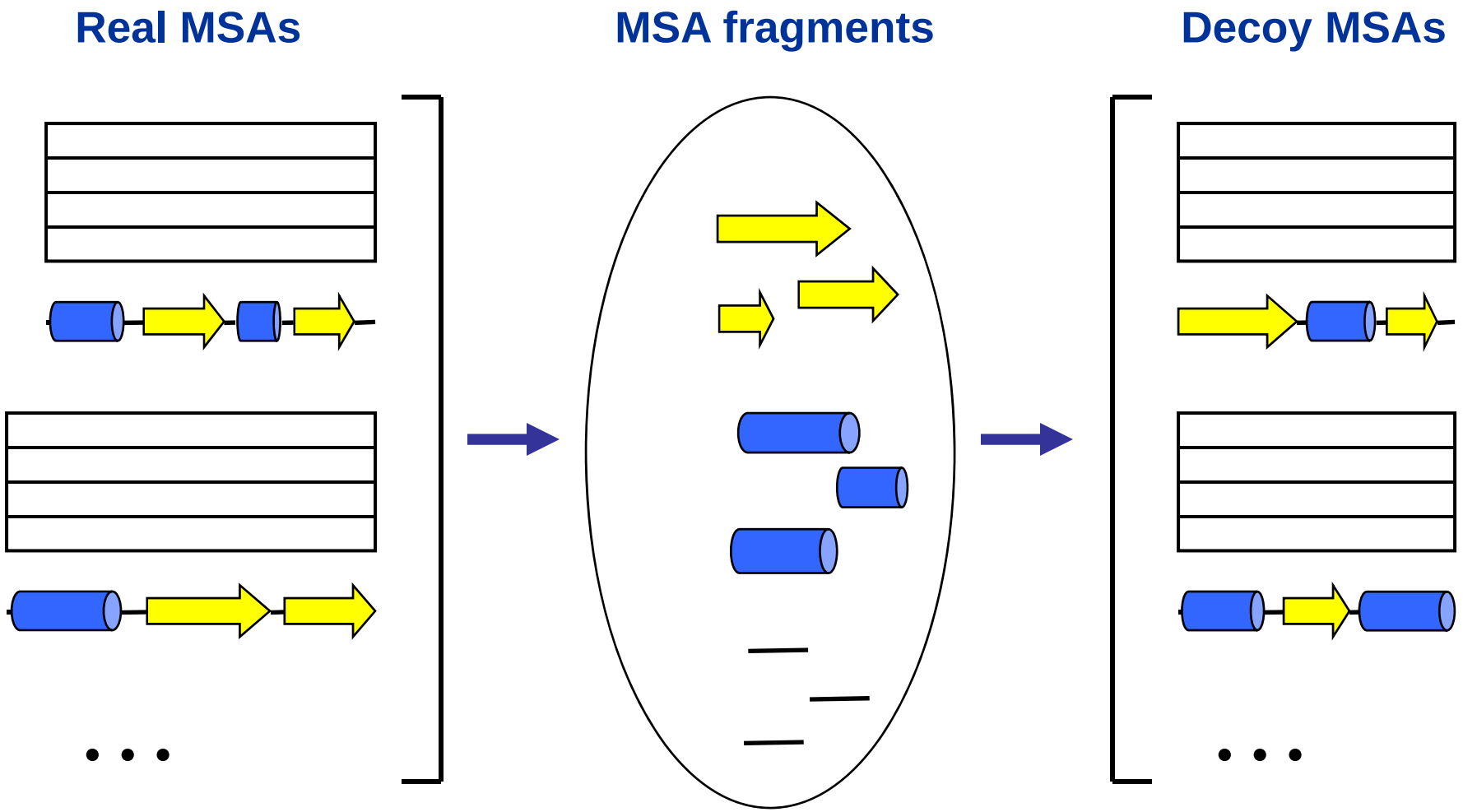
Independent positions: shuffling positions makes decoy alignments

This model works very well in BLAST and PSI-BLAST,

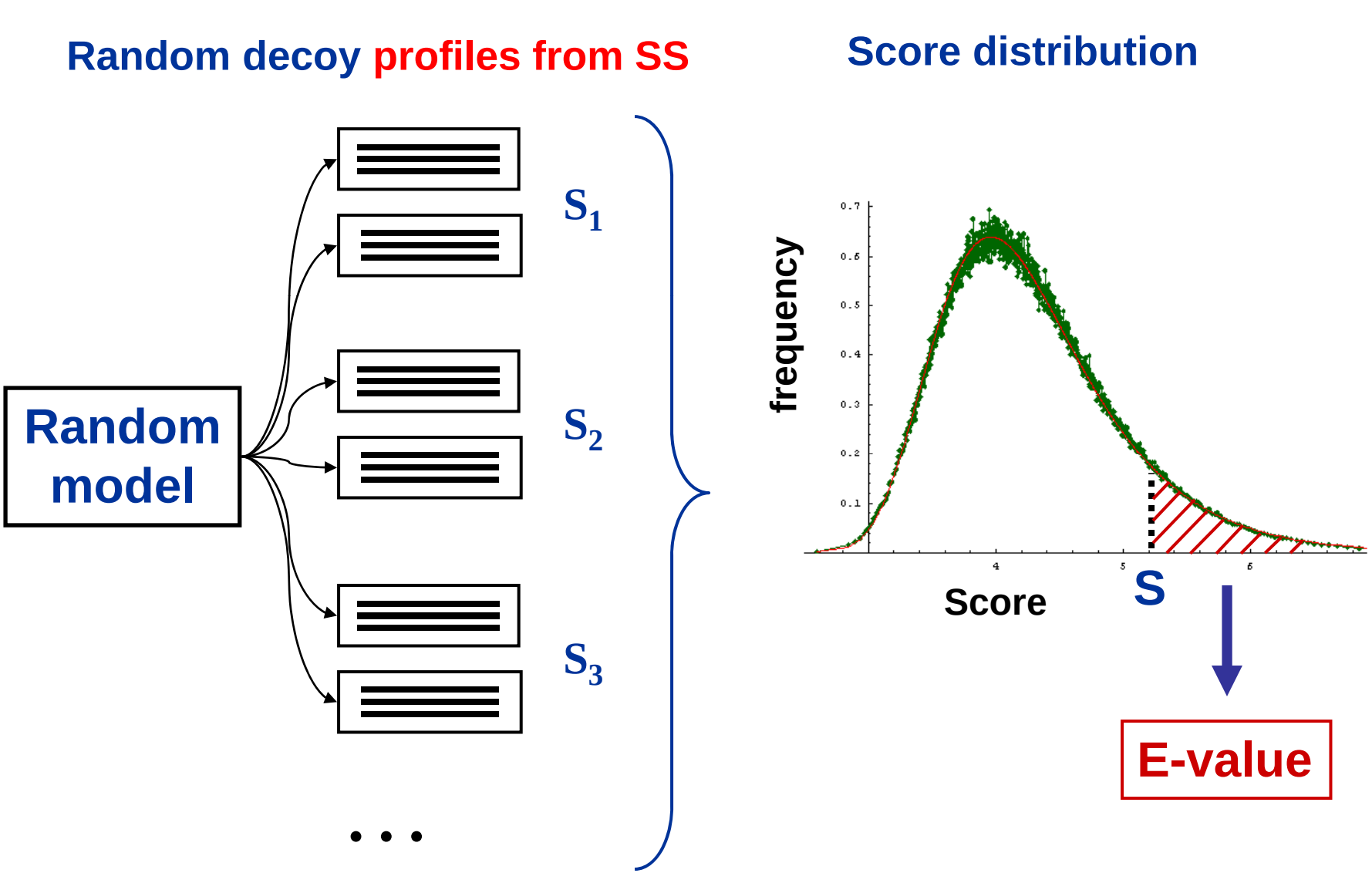
however, maybe more realistic models work better

Reproducing protein features:

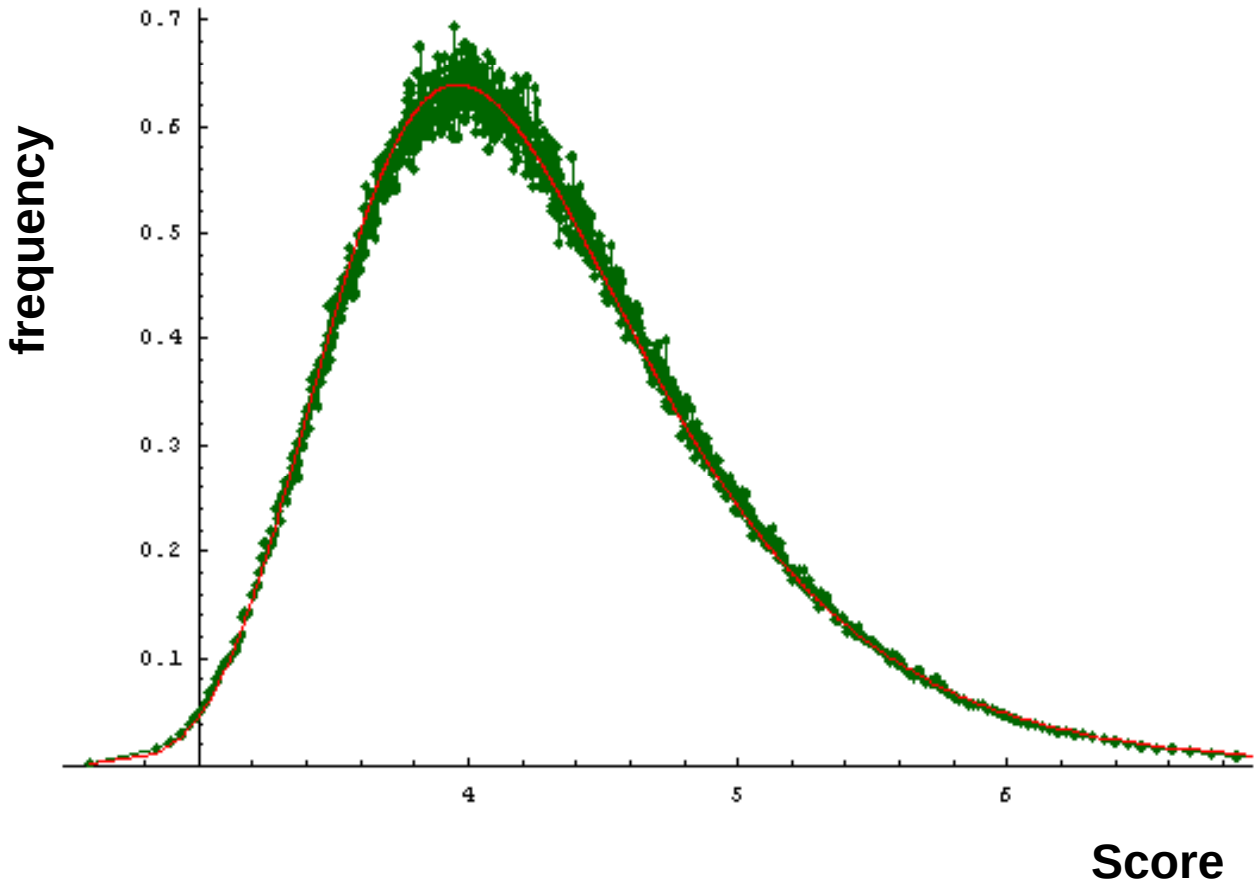
Real secondary structure elements are used as building blocks for decoy MSA



Estimates of statistical significance are based on a random model of alignment comparison



Distribution of scores for random MSA comparison



Describe empirical distribution with a continuous density function

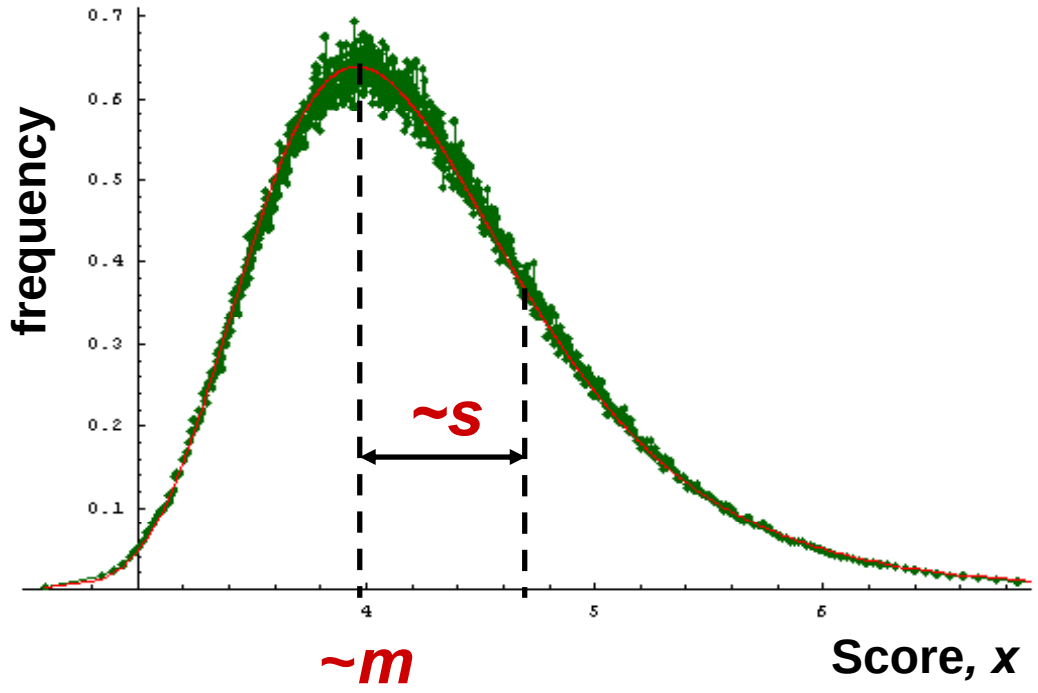
Gumbel Extreme Value Distribution (EVD) is traditionally used to describe similarity cores

EVD pdf:

$$f(x) = C_1 \exp(-e^{-\frac{x-m}{s}} - \frac{x-m}{s})$$

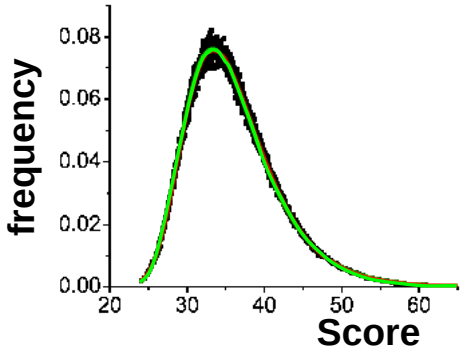
***m*: location parameter**

***s*: scale parameter**



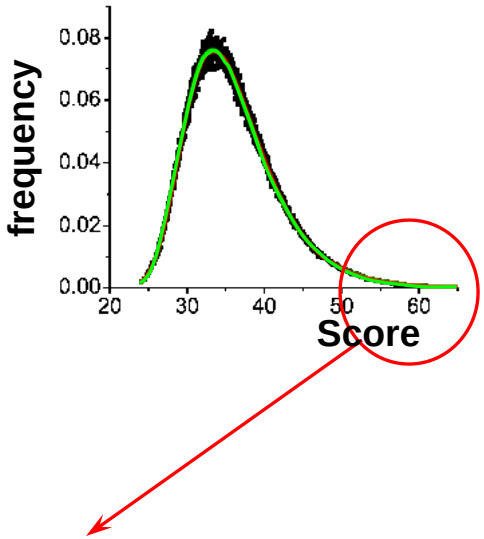
EVD does not fit empirical score distributions

$$f(x) = C_1 \exp(-e^{-\frac{x-m}{s}} - \frac{x-m}{s})$$



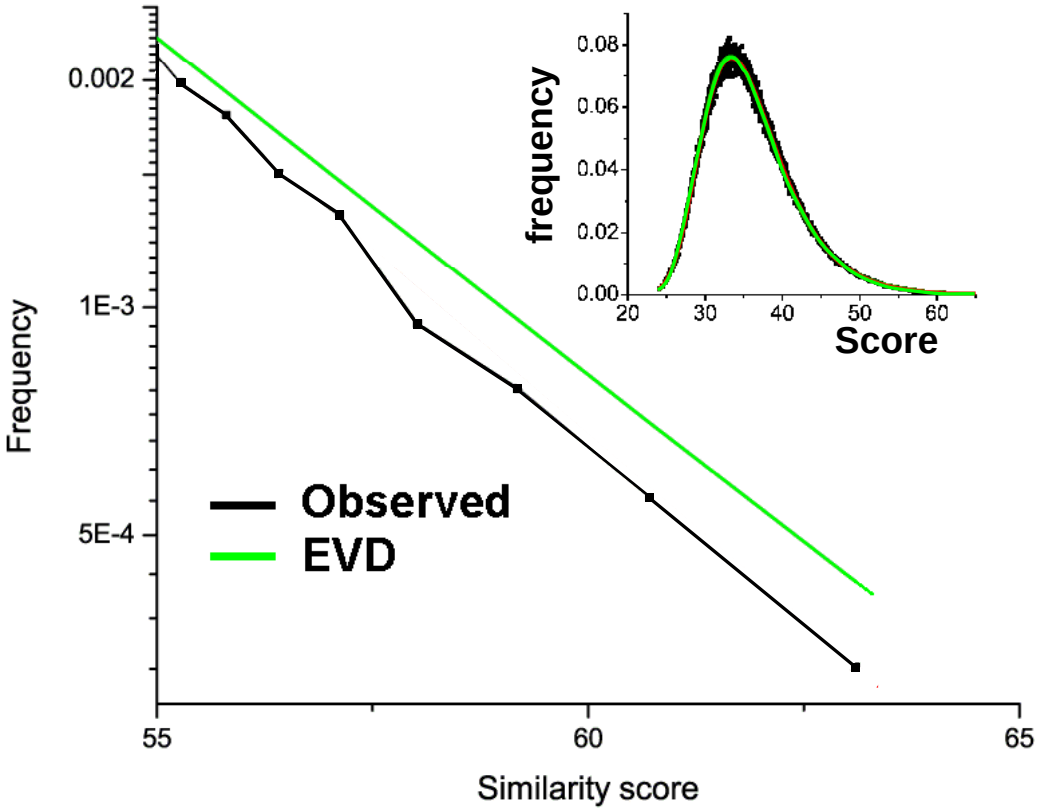
EVD does not fit empirical score distributions

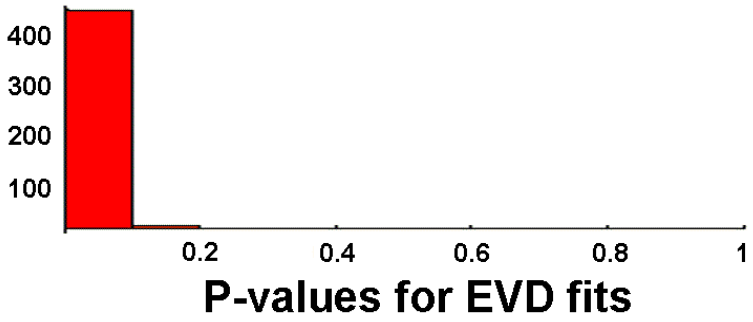
$$f(x) = C_1 \exp(-e^{-\frac{x-m}{s}} - \frac{x-m}{s})$$



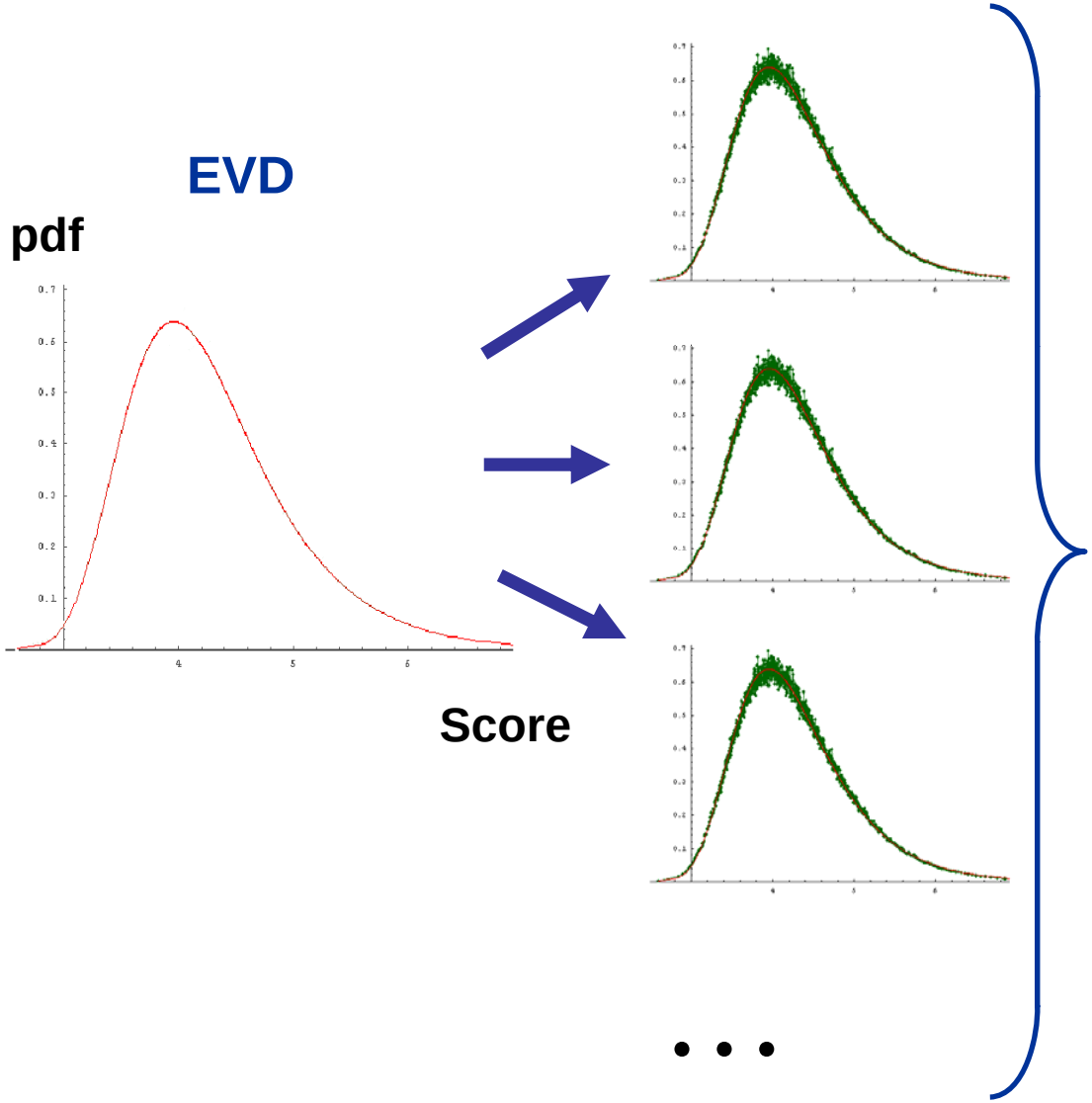
EVD does not fit empirical score distributions

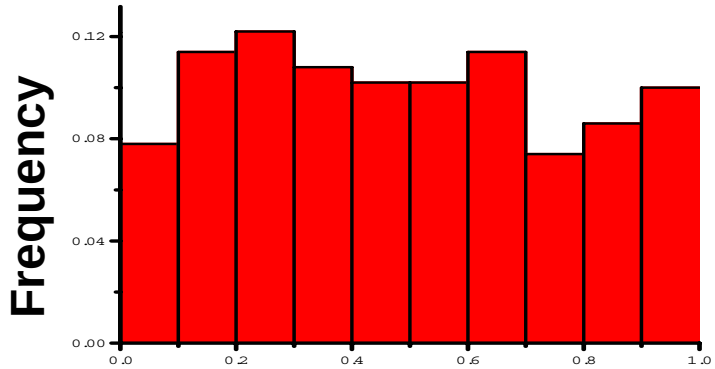
$$f(x) = C_1 \exp(-e^{-\frac{x-m}{s}} - \frac{x-m}{s})$$





For data generated from the same distribution, fitting P-values are distributed uniformly





P-values for EVD fits

Scores generated by SS-based model do not obey other standard statistical distributions

Distributions of Pearson system

Distributions of Johnson system

Inverse Gaussian (Wald) distribution

Burr

Weibul

Tukey (lambda)

Non-central chi square

Non-central t

χ^2 goodness-of-fit

does not pass

P-values <~ 10^{-5}

We had to invent a new distribution

How?

Modify EVD!

EVD pdf: $$f(x) = C_1 \exp(-e^{-\frac{x-m}{s}} - \frac{x-m}{s})$$

Power EVD pdf: $$f(x) = C_2 \exp(-e^{-\frac{x-m}{s}} - \beta \frac{x^{\alpha} - m^{\alpha}}{s^{\alpha}})$$

WOW!

A new distribution, power EVD (PEVD), is created by modification of EVD

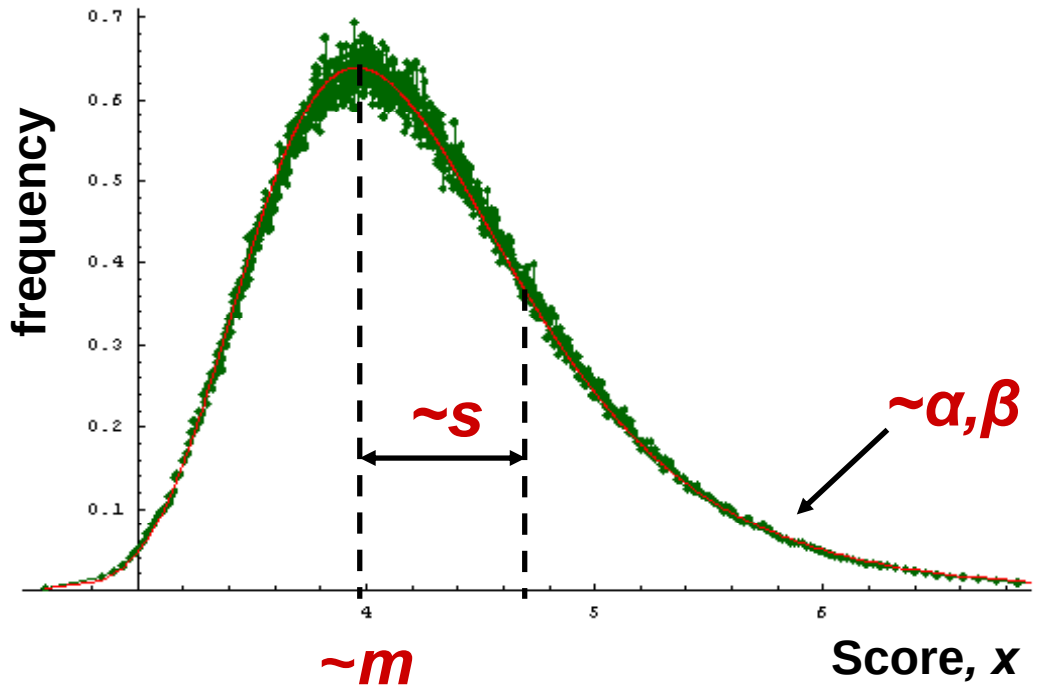
EVD pdf: $$f(x) = C_1 \exp(-e^{-\frac{x-m}{s}} - \frac{x-m}{s})$$

PEVD pdf: $$f(x) = C_2 \exp(-e^{-\frac{x-m}{s}} - \beta \frac{x^{\alpha} - m^{\alpha}}{s^{\alpha}})$$

***m*: location parameter**

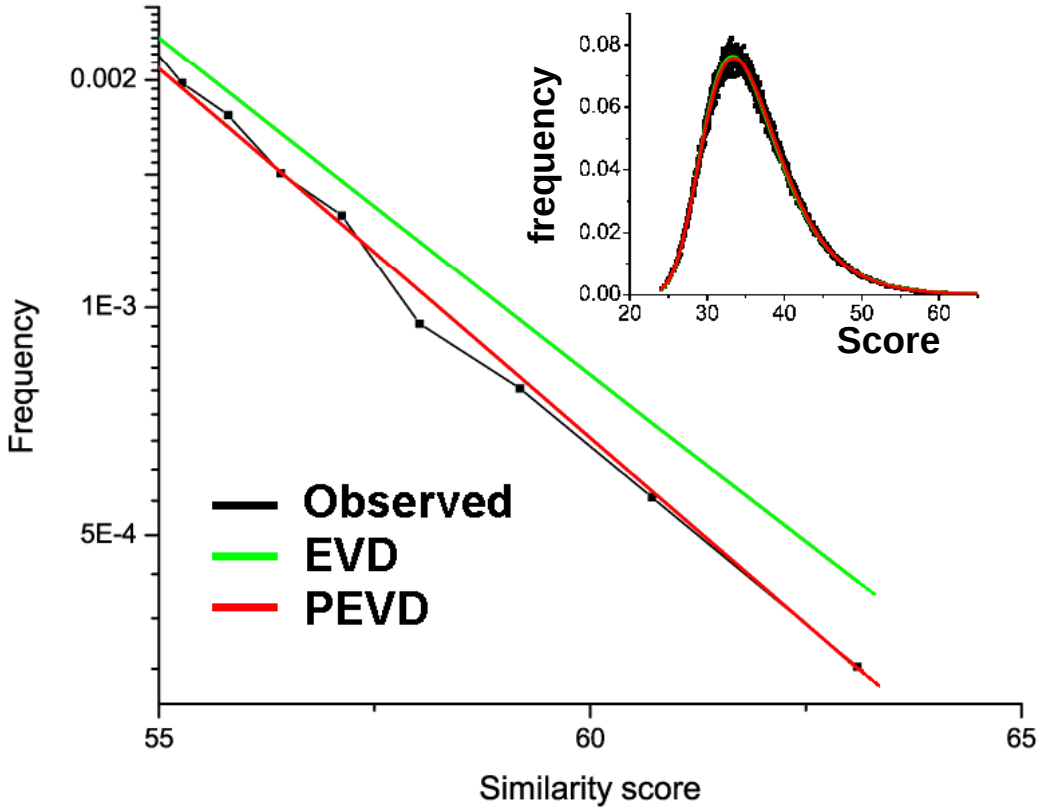
s: scale parameter

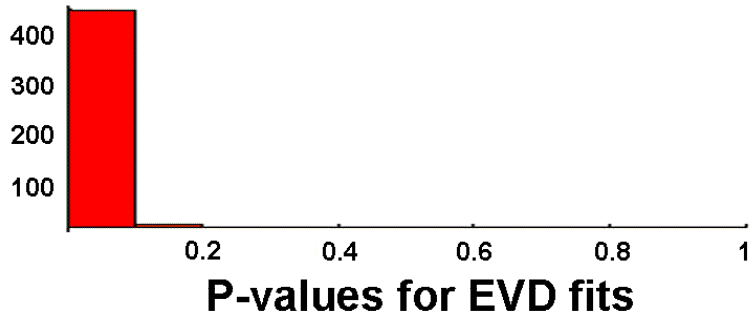
α, β: shape parameters

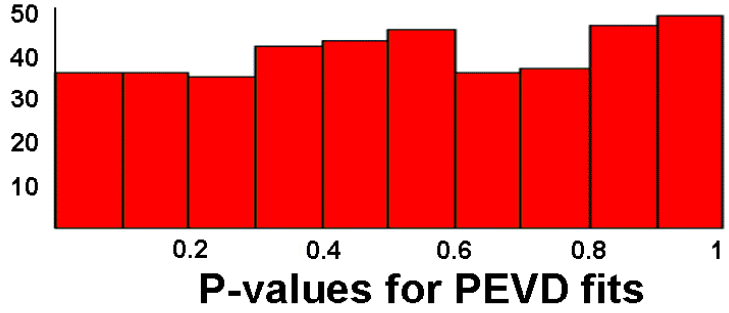


Power EVD precisely fits empirical score distributions

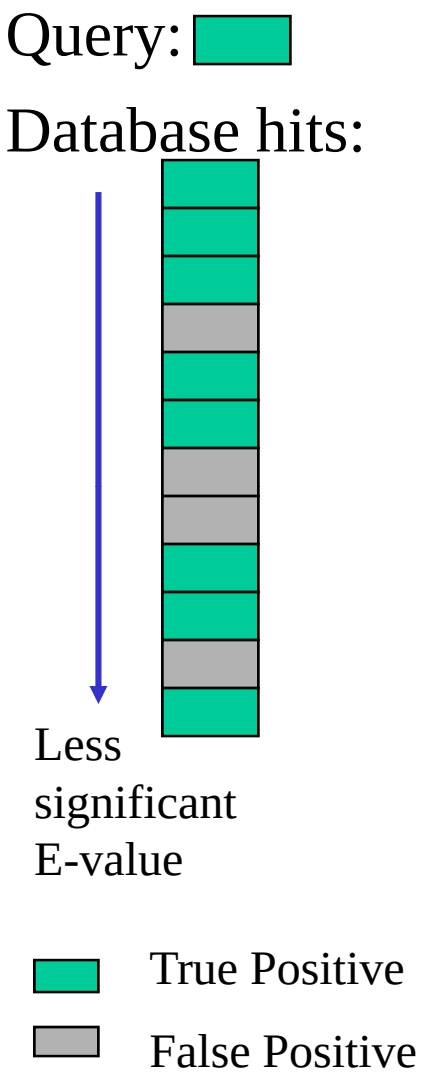
$$f(x) = C_2 \exp(-e^{-\frac{x-m}{s}} - \beta \frac{x^\alpha - m^\alpha}{s^\alpha})$$



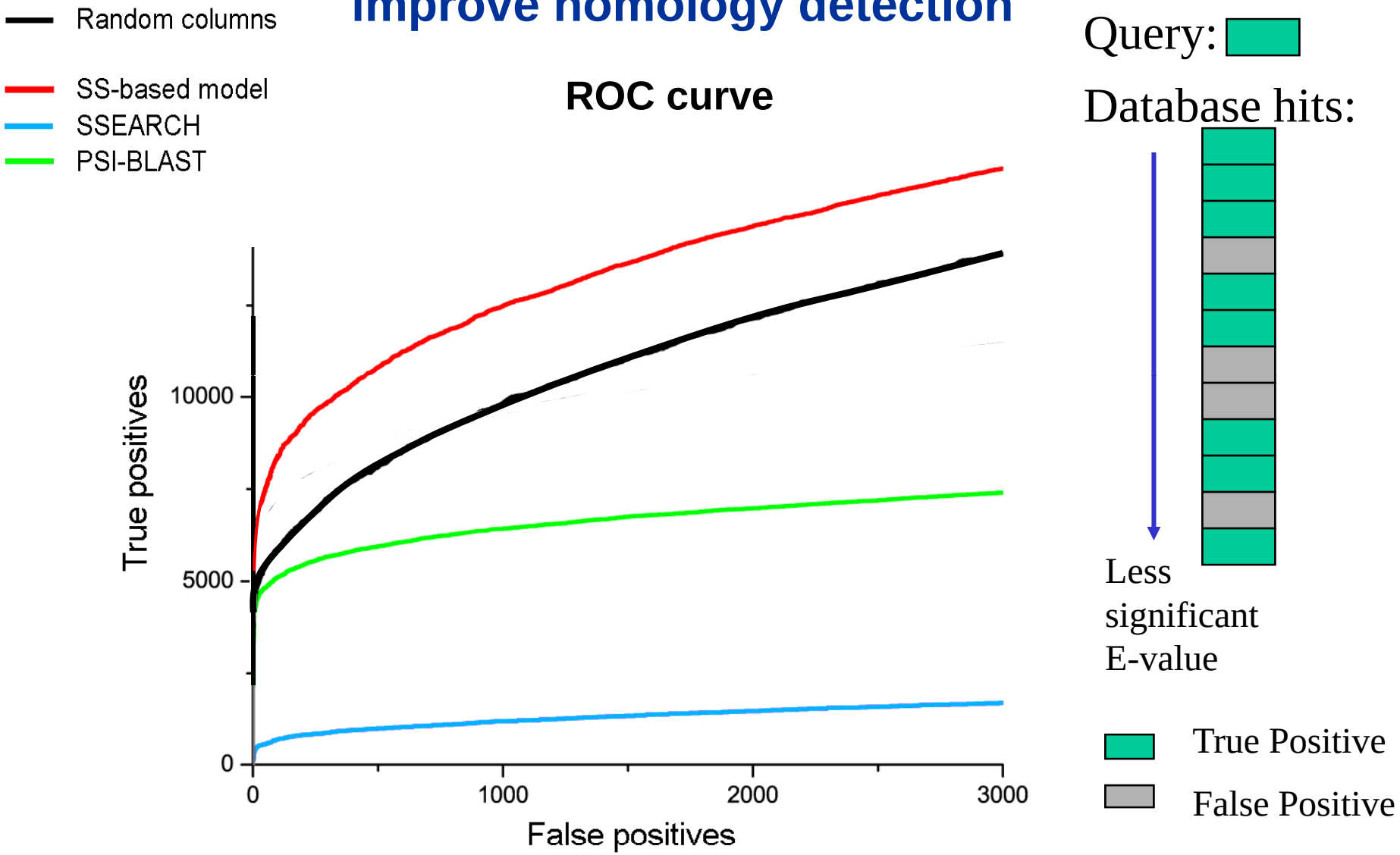




The new random model + new distribution improve homology detection



The new random model + new distribution improve homology detection



Benchmark: 2900 PSI-BLAST alignments for SCOP domain representatives with known relationships

Summary

- We developed a realistic random model that simulates random MSA comparison by mimicking native protein secondary structure

- We developed a precise analytical approximation of the simulated score distributions, based on a new distribution function, PEVD

- Applied to protein similarity searches, the new model produces more realistic E-values and (unexpectedly) improves homology detection

2. Towards accurate multiple sequence alignments of distantly related proteins

Jimin Pei

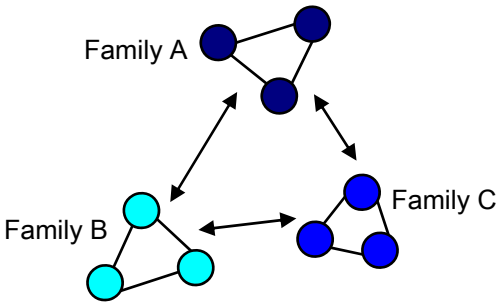
Multiple sequence alignment

```
BSUB00  RMAHYDSLTDLPNRRHAISHLTKVLNREHSLHYNTVVFFLDLNRFKVINDAL
ECU738  VMSTRDGMTGVYNRRHWETMLRNEFDNCRRHNRDATLLIIDIDHFKSINDTW
D90790  HEVGMDVLTKLLNRRFLPTIFKREIAHANRTGTPLSVLIIDVDKFKEINDTW
SYCSLL  QISSLDALTQVGNRYLFDSTLEREWQRLQRIREPLALLLCDVDFFKGFNDNY
ECAE00  NIAHRDPLTNIFNRNYFFNEL--TVQSASAQKTPYCVMIMDIDHFKKVNDTW
AF0348  QAANVDSLTGLANRAAYNAHM-ERLTAADAPS--IGLLLIDVDRLKQVNDIL
D90796  IRSNMDVLTGLPGRRVLDESFDHQLRNAEPLN--LYLMLLDIDRFKLVNDTY
Y4LL_R  HMARHDALTGLPNRQFLREEF-ERLSDHIAPSTRLAILCLDLDGFKAINDAY
Y07I_M  YLADHDDLTGLHNRRALLQHLDQRLAPGQPGP--VAALFLDLDRLKAINDYL
......
```

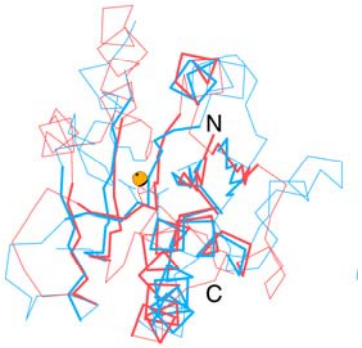
Multiple sequence alignment

```
BSUB00   RMAHYDSLTDLPNRRHAISHLTKVLNREHSLHYNTVVFFLDLNRFKVINDAL
ECU738   VMSTRDGMTGVYNRRHWETMLRNEFDNCRRHNRDATLLIIDIDHFKSINDTW
D90790   HEVGMDVLTKLLNRRFLPTIFKREIAHANRTGTPLSVLIIDVDKFKEINDTW
SYCSLL   QISSLDALTQVGNRYLFDSTLEREWQRLQRIREPLALLLCDVDFFKGFNDNY
ECAE00   NIAHRDPLTNIFNRNYFFNEL--TVQSASAQKTPYCVMIMDIDHFKKVNDTW
AF0348   QAANVDSLTGLANRAAYNAHM-ERLTAADAPS--IGLLLIDVDRLKQVNDIL
D90796   IRSNMDVLTGLPGRRVLDESFDHQLRNAEPLN--LYLMLLDIDRFKLVNDTY
Y4LL_R   HMARHDALTGLPNRQFLREEF-ERLSDHIAPSTRLAILCLDLDGFKAINDAY
Y07I_M   YLADHDDLTGLHNRRALLQHLDQRLAPGQPGP--VAALFLDLDRLKAINDYL
.....
```

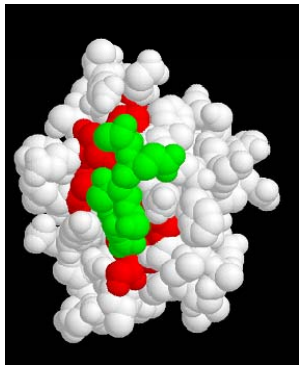
Protein similarity search and classification



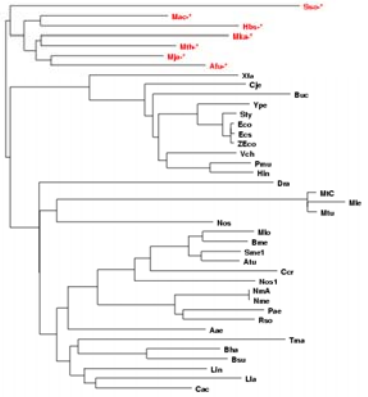
Structure modeling



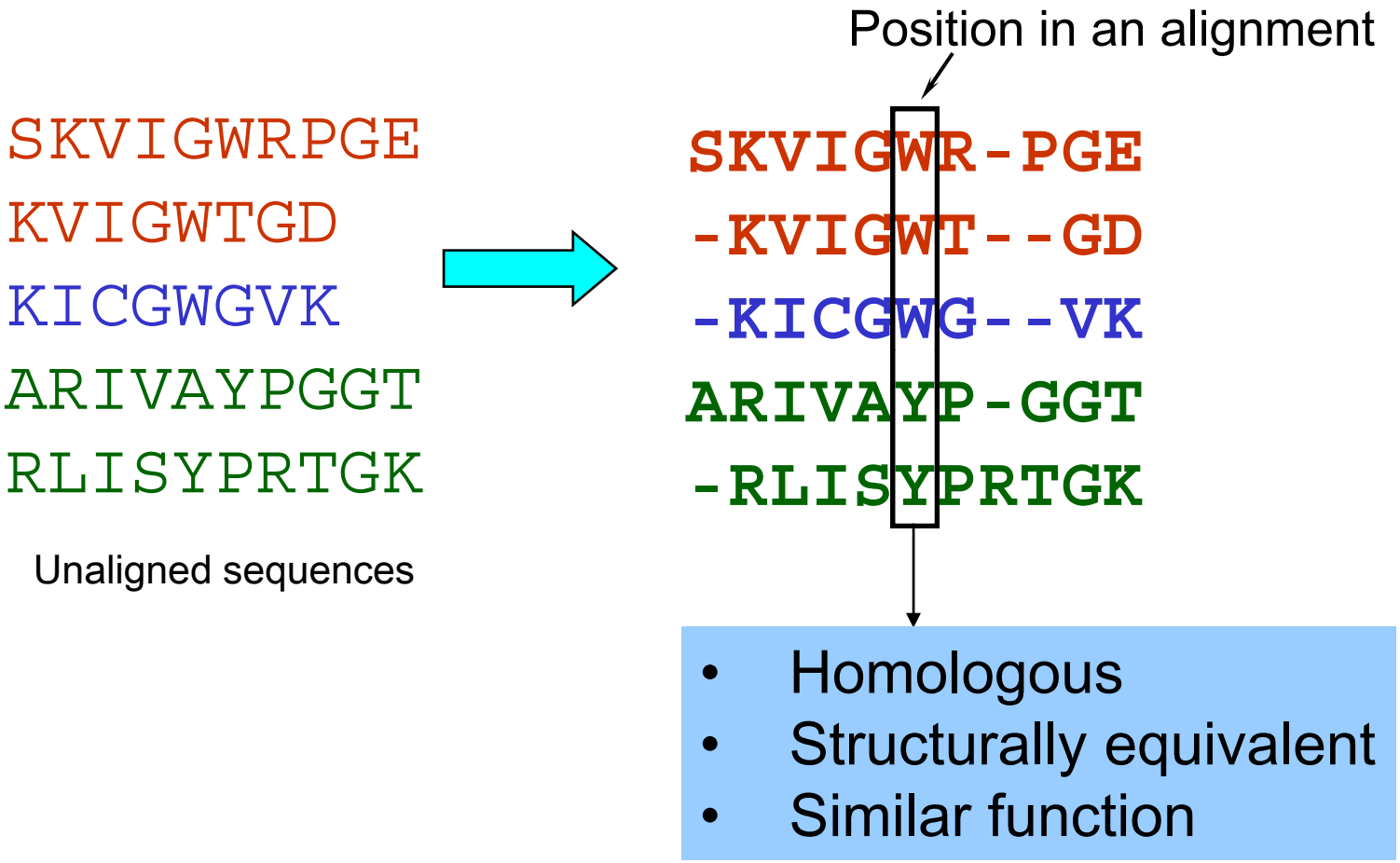
Active site prediction experimental design



Phylogenetic analysis



Meaning of alignments

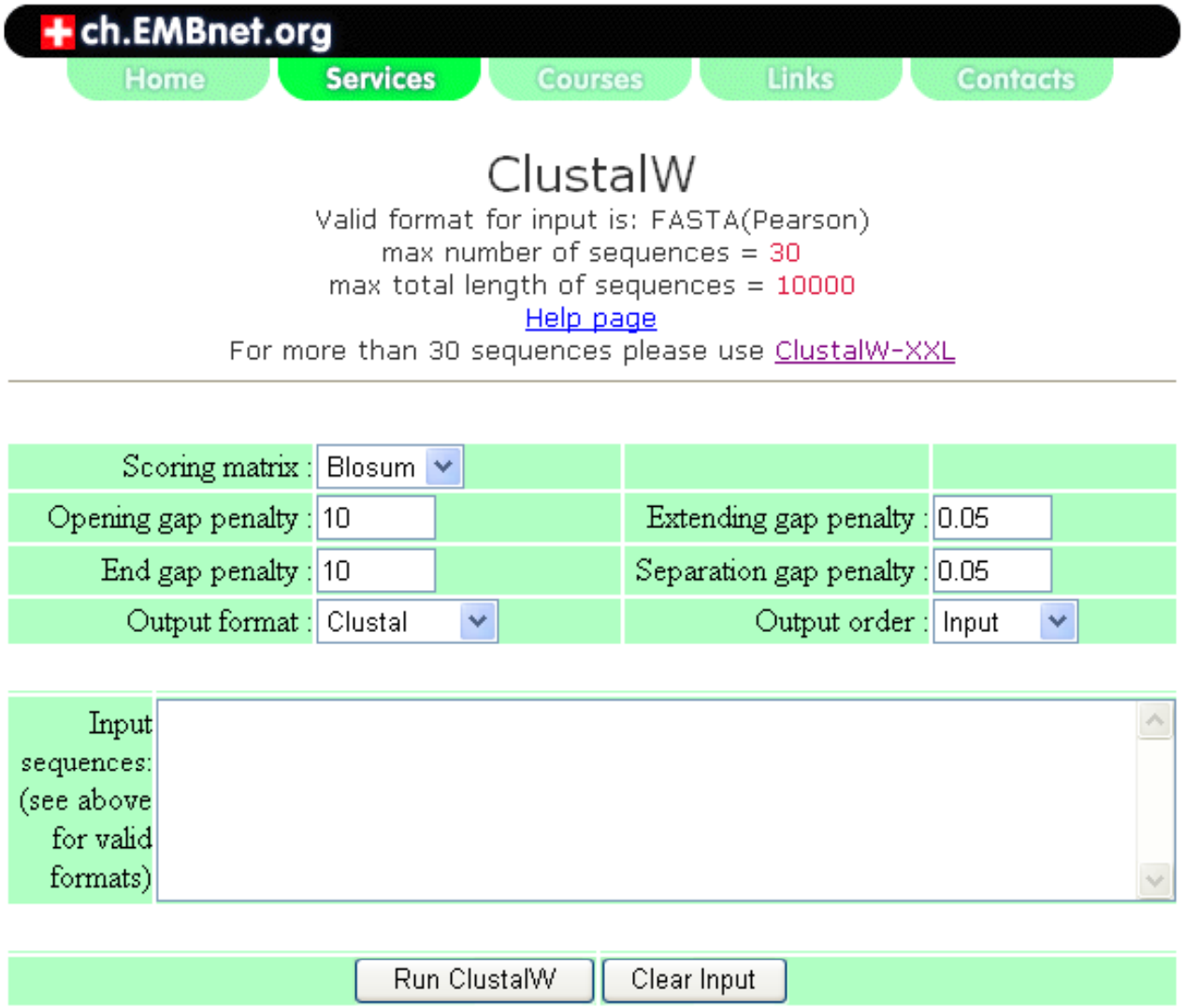


How is the alignment made?

ClustalW –

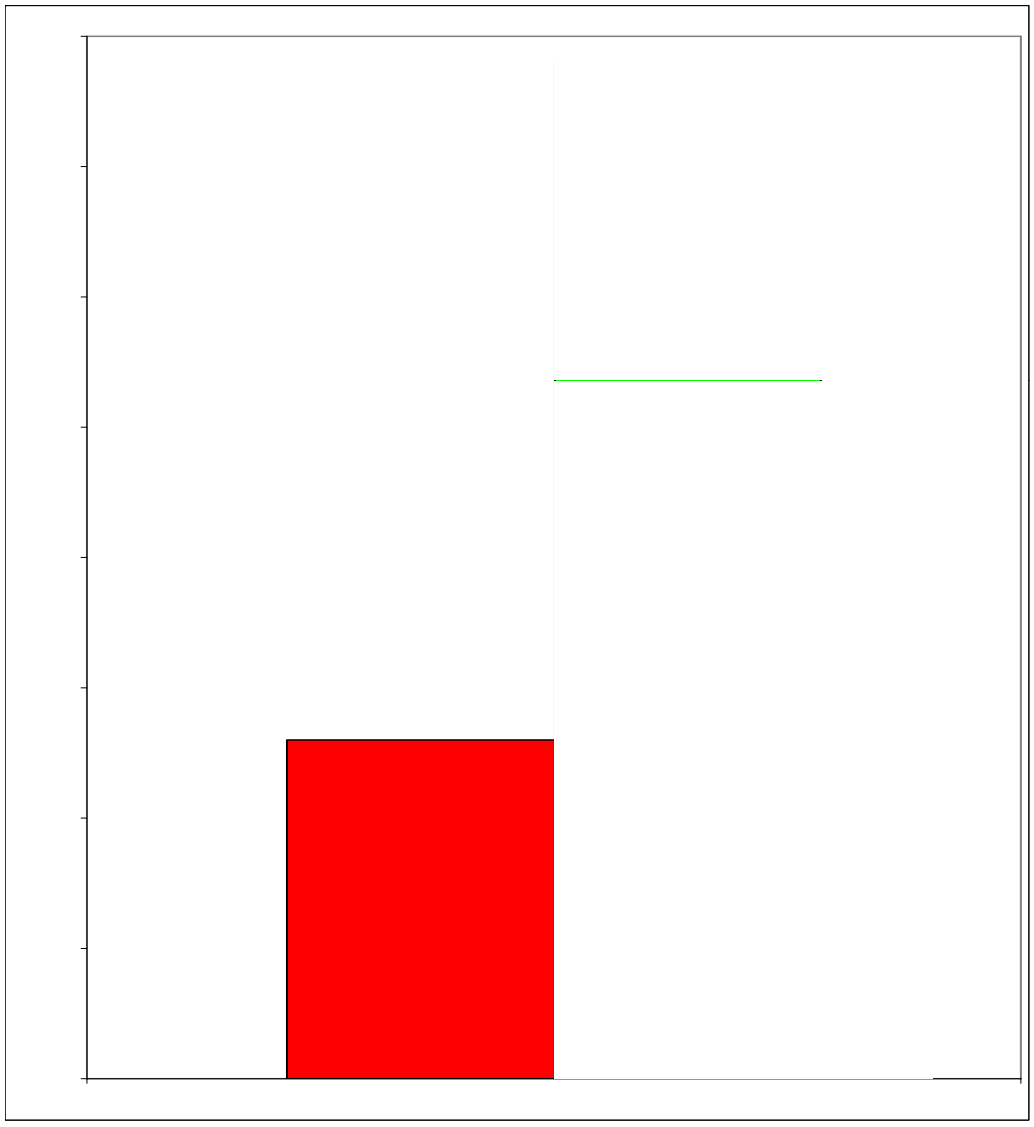
the most widely used alignment program

ClustalW – the most widely used program



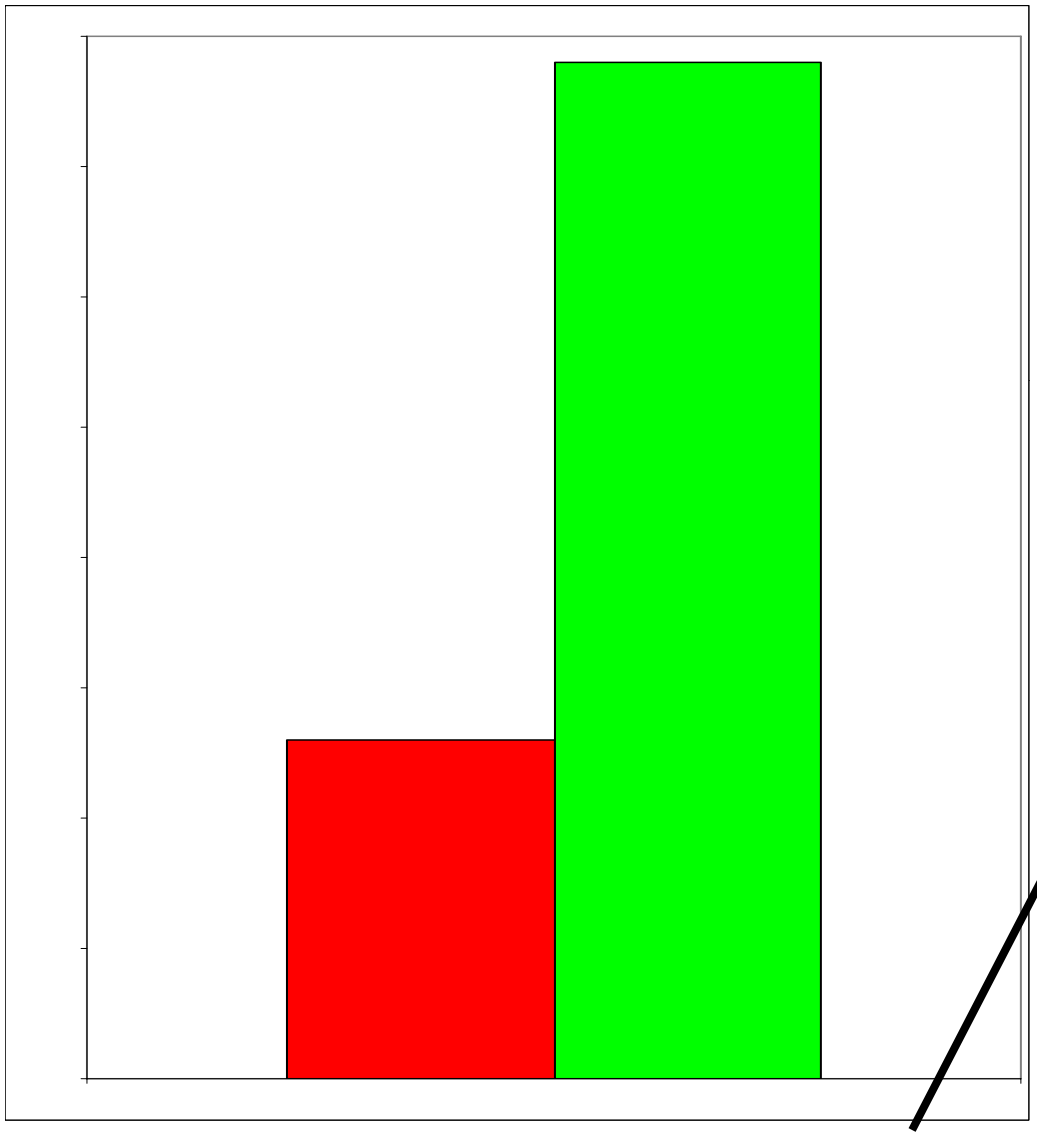
Thompson et al. (1994). http://www.ch.embnet.org/software/ClustalW.html

How accurate are these alignments?



ClustalW accuracy

How accurate are these alignments?

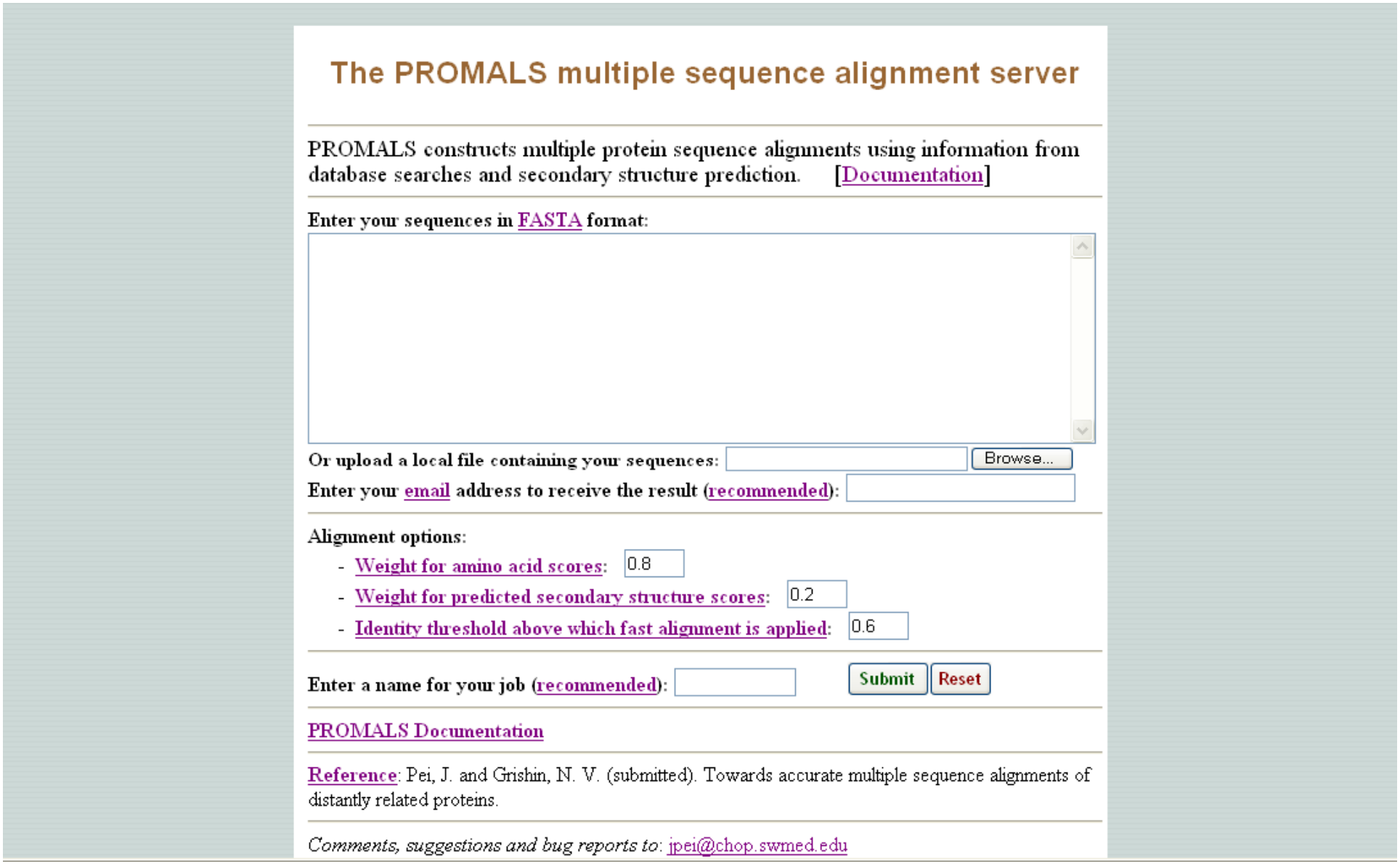


About 3 times better than ClustalW

PROMALS:

(**PRO**file **M**ultiple **A**lignment with predicted **L**ocal **S**tructure)

http://prodata.swmed.edu/promals



The PROMALS multiple sequence alignment server

PROMALS constructs multiple protein sequence alignments using information from database searches and secondary structure prediction. [Documentation]

Enter your sequences in FASTA format:

Or upload a local file containing your sequences: Browse...

Enter your email address to receive the result (recommended):

Alignment options:

- Weight for amino acid scores: 0.8
- Weight for predicted secondary structure scores: 0.2
- Identity threshold above which fast alignment is applied: 0.6

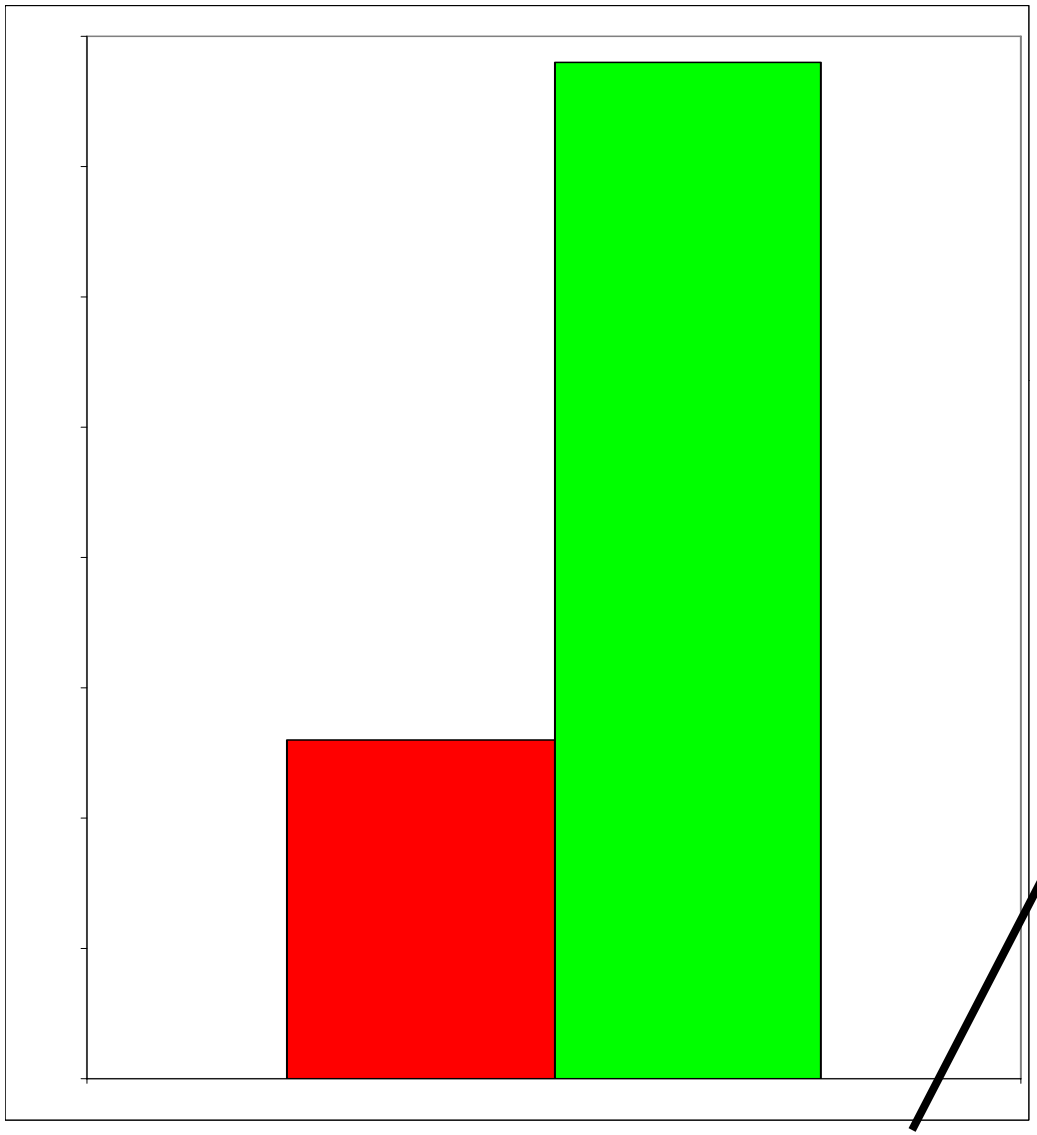
Enter a name for your job (recommended): Submit Reset

PROMALS Documentation

Reference: Pei, J. and Grishin, N. V. (submitted). Towards accurate multiple sequence alignments of distantly related proteins.

Comments, suggestions and bug reports to: jpei@chop.swmed.edu

What did we do to achieve this?



First of all,

ClustalW is not that bad …

… for similar sequences

```
Q2BMK3    MLAKTVREQIQRPADLVARYGGEEFIVVLPDTDEEGAMAVAGQICVAVAS
Q3A8D4    QVARMLQSVVARPGDLVARYGGEEFALILPQTD-HGAKFLGESCRAAVAG
Q36PG9    ALASILSDEVQRSGDLVARYGGEEFAILLPTTDVAGAQQVAERMRLSVAR
Q2BQL8    TVAQTIKHSIQRAQDMVCRYGGEEFVVILPETDLDGAQMIAERIRKAIAK
Q3XUK3    ALAHTISL-HLRPGDIAARYGGEEFAVVLPDTDAVSGRMIAERLRTAVEA
Q9HXT9    QVAGAIREGCSRSSDLAARYGGEEFAMVLPGTSPGGARLLAEKVRRTVES
P73713    TIGRILQSNIRGS-DIACRYGGEEMTIVLPQTSLEDTLVKAESLRQAIAS
Q36SI5    MVGDVLATCFRGS-DTVCRYGGEEFSVLMPGASLDEARQRAEQLRAAISA
Q747B7    EAAAVFRGCIRTS-DIAARYGGEEFVVIMPETTRELALLAAEKLRRAVEE
Q2DK38    KTADIIKASLRDM-DIVARYGGEEFCAILPGTSKKESIVVAERIRVGIEK
```

ClustalW good alignment

… for similar sequences

ClustalW good alignment

Here are distantly related sequences: diguanylate cyclase and adenylate cyclase

ClustalW alignment

```
1w25 ------NRRYMTGQLDSLVKRATLGGDPVSALL-------------IDIDFFKKINDTFGHDIGDEV-------LREFALRLAS
1wc4 -PEPRLITILFSDIVGFTRMSNALQSQGVAELLNEYLGEMTRAVFENQGTVDKFVGDAIMALYGAPEEMSPSEQVRRAIATARQ

1w25 NVRAI-DLPCRYGGEE-----------FVVIMPDTALADALRI-AERIRMHVSGSPFTVAHGREML--NVTISIGVSATAGEGD
1wc4 MLVALEKLNQGWQERGLVGRNEVPPVRFRCGIHQGMAVVGLFGSQERSDFTAIGPSVNIAARLQEATAPNSIMVSAMVAQYVPD

1w25 TPEALLKRADEGVYQAKASGRNAVVGKAA--
1wc4 E-----EIIKREFLELKGIDEPVMTCVINPN
```

sequence identity = 12%

DALI alignment based on structural comparison

```
1w25 NRRYMTGQLDSLVKRATLGGDPVSALLIDIDFFKKINDTFGHDIGDEVLREFALRLASNVRA-IDLP-CRYGGEEFVVIMPDT-
1wc4 -------------PEPR----LITILFSDIVGFTRMSNALQSQGVAELLNEYLGEMTRAVFENQGTVDKFVG-DAIMALYGAPE

1w25 ------ALADALRIAERIRMHVSG-SPFTVAHGREMLN------VTISIGVSATAGEGDT----------PEALLKRADEGVYQ
1wc4 EMSPSEQVRRAIATARQMLVALEKLNQGW-QERGLVGRNEVPPVRFRCGIHQGMAVVGLFGSQERSDFTAIGPSVNIAARLQEA

1w25 AKASGRNAVVGKAA---------------------------------
1wc4 TA---PNSIMVSAMVAQYVPDEEIIKREFLELKGIDEPVMTCVINPN
```

sequence identity = 12%

1. Pei and Grishin 2001 2. Steegborn et al. 2005 3.Holm and Sander 1998

Red: alpha-helix blue: beta-strand

1. ClustalW alignment

```
1w25 ------NRRYMTGQLDSLVKRATLGGDPVSALL------------IDIDFFKKINDTFGHDIGDEV-------LREFALRLAS
1wc4 -PEPRLITILFSDIVGFTRMSNALQSQGVAELLNEYLGEMTRAVFENQGTVDKFVGDAIMALYGAPEEMSPSEQVRRAIATARQ

1w25 NVRAI-DLPCRYGGEE----------FVVIMPDTALADALRI-AERIRMHVSGSPFTVAHGREML--NVTISIGVSATAGEGD
1wc4 MLVALEKLNQGWQERGLVGRNEVPPVRFRCGIHQGMAVVGLFGSQERSDFTAIGPSVNIAARLQEATAPNSIMVSAMVAQYVPD

1w25 TPEALLKRADEGVYQAKASGRNAVVGKAA--
1wc4 E-----EIIKREFLELKGIDEPVMTCVINPN
```

: α-helix aligned to β-strand!

2. DALI alignment based on structural comparison

```
1w25 NRRYMTGQLDSLVKRATLGGDPVSALLIDIDFFKKINDTFGHDIGDEVLREFALRLASNVRA-IDLP-CRYGGEEFVVIMPDT-
1wc4 ------------PEPR----LITILFSDIVGFTRMSNALQSQGVAELLNEYLGEMTRAVFENQGTVDKFVG-DAIMALYGAPE

1w25 ------ALADALRIAERIRMHVSG-SPFTVAHGREMLN------VTISIGVSATAGEGDT----------PEALLKRADEGVYQ
1wc4 EMSPSEQVRRAIATARQMLVALEKLNQGW-QERGLVGRNEVPPVRFRCGIHQGMAVVGLFGSQERSDFTAIGPSVNIAARLQEA

1w25 AKASGRNAVVGKAA--------------------------------
1wc4 TA---PNSIMVSAMVAQYVPDEEIIKREFLELKGIDEPVMTCVINPN
```

Accuracy of the above ClustalW alignment: **0%**

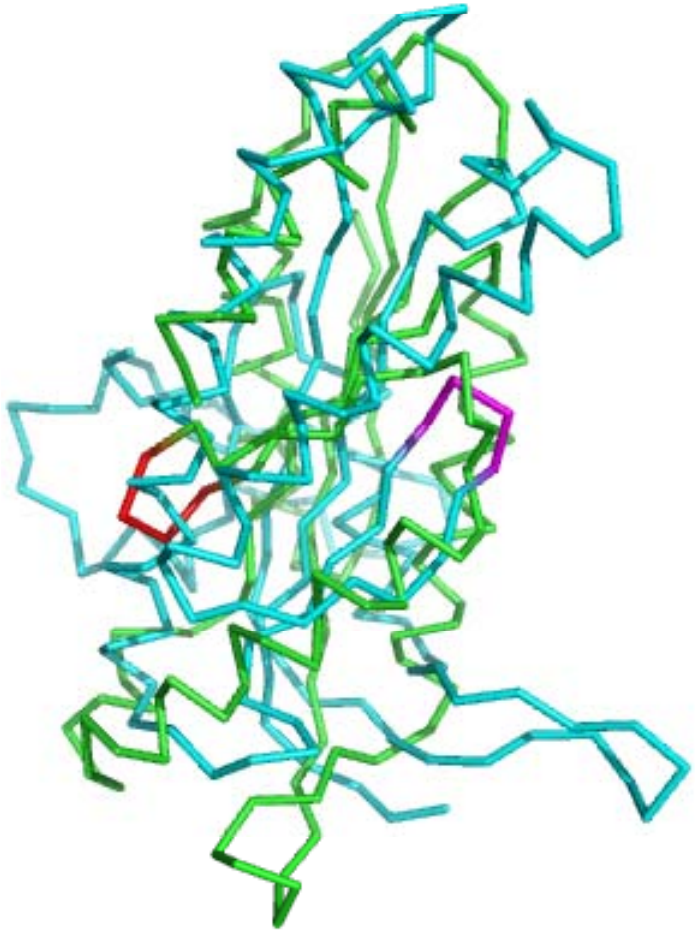
Alignment-based structural superposition

1. ClustalW alignment

```
1w25 ------NRRYMTGQLDSLVKRATLGGDPVSALL-------------IDIDFFKKINDTFGHDIGDEV-------LREFALRLAS
1wc4 -PEPRLITILFSDIVGFTRMSNALQSQGVAELLNEYLGEMTRAVFENQGTVDKFVGDAIMALYGAPEEMSPSEQVRRAIATARQ

1w25 NVRAI-DLPCRYGGEE-----------FVVIMPDTALADALRI-AERIRMHVSGSPFTVAHGREML--NVTISIGVSATAGEGD
1wc4 MLVALEKLNQGWQERGLVGRNEVPPVRFRCGIHQGMAVVGLFGSQERSDFTAIGPSVNIAARLQEATAPNSIMVSAMVAQYVPD

1w25 TPEALLKRADEGVYQAKASGRNAVVGKAA--
1wc4 E-----EIIKREFLELKGIDEPVMTCVINPN
```

ClustalW superposition

Alignment-based structural superposition



ClustalW superposition

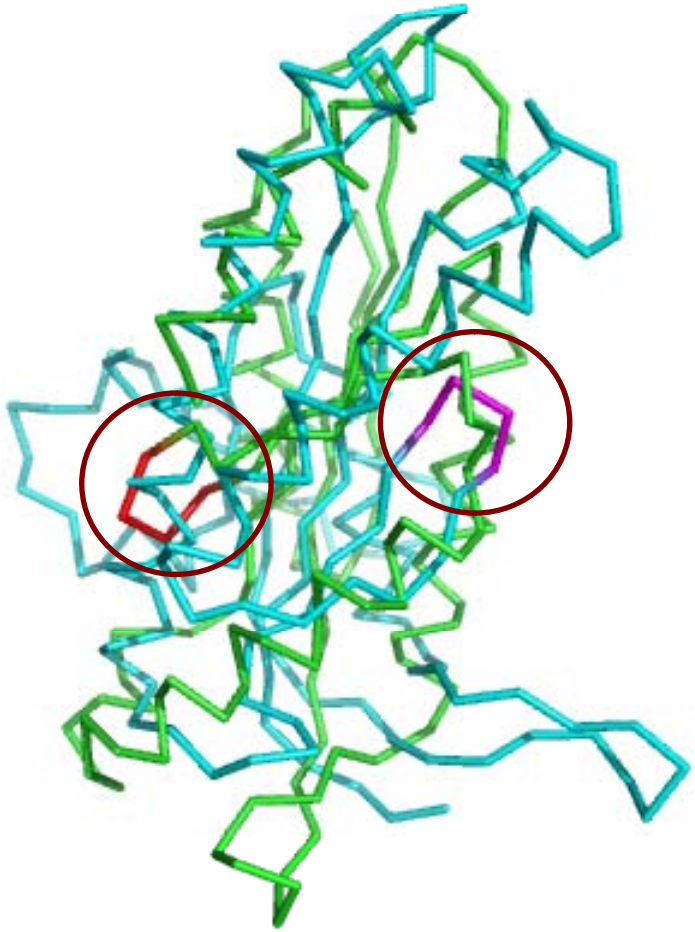
Alignment-based structural superposition

```
1w25 NRRYMTGQLDSLVKRATLGGDPVSALLIDIDFFKKINDTFGHDIGDEVLREFALRLASNVRA-IDLP-CRYGGEEFVVIMPDT-
1wc4 -------------PEPR----LITILFSDIVGFTRMSNALQSQGVAELLNEYLGEMTRAVFENQGTVDKFVG-DAIMALYGAPE

1w25 ------ALADALRIAERIRMHVSG-SPFTVAHGREMLN------VTISIGVSATAGEGDT----------PEALLKRADEGVYQ
1wc4 EMSPSEQVRRAIATARQMLVALEKLNQGW-QERGLVGRNEVPPVRFRCGIHQGMAVVGLFGSQERSDFTAIGPSVNIAARLQEA

1w25 AKASGRNAVVGKAA--------------------------------
1wc4 TA---PNSIMVSAMVAQYVPDEEIIKREFLELKGIDEPVMTCVINPN
```

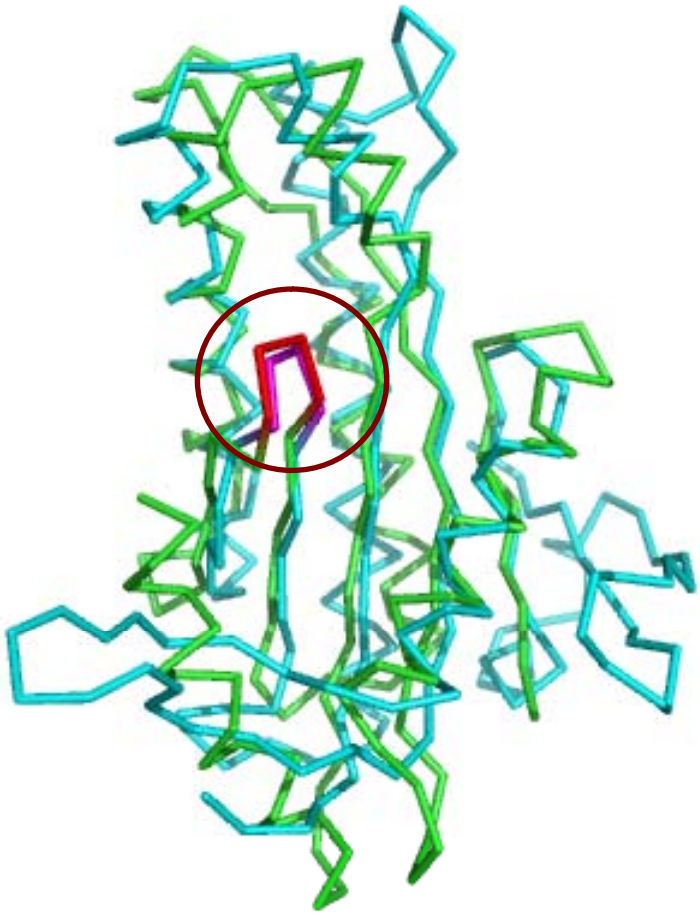
ClustalW superposition

DALI superposition

Alignment-based structural superposition

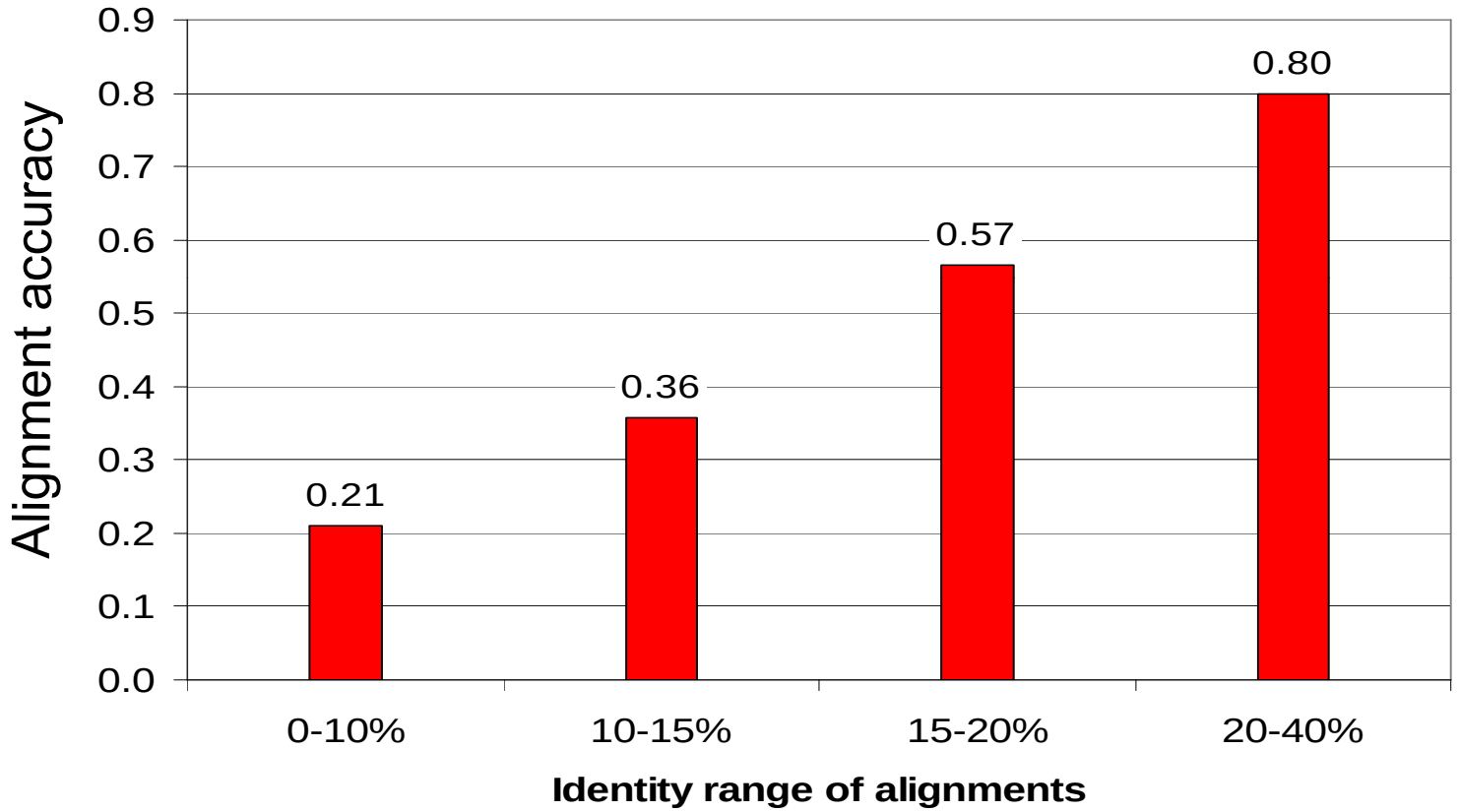


ClustalW superposition



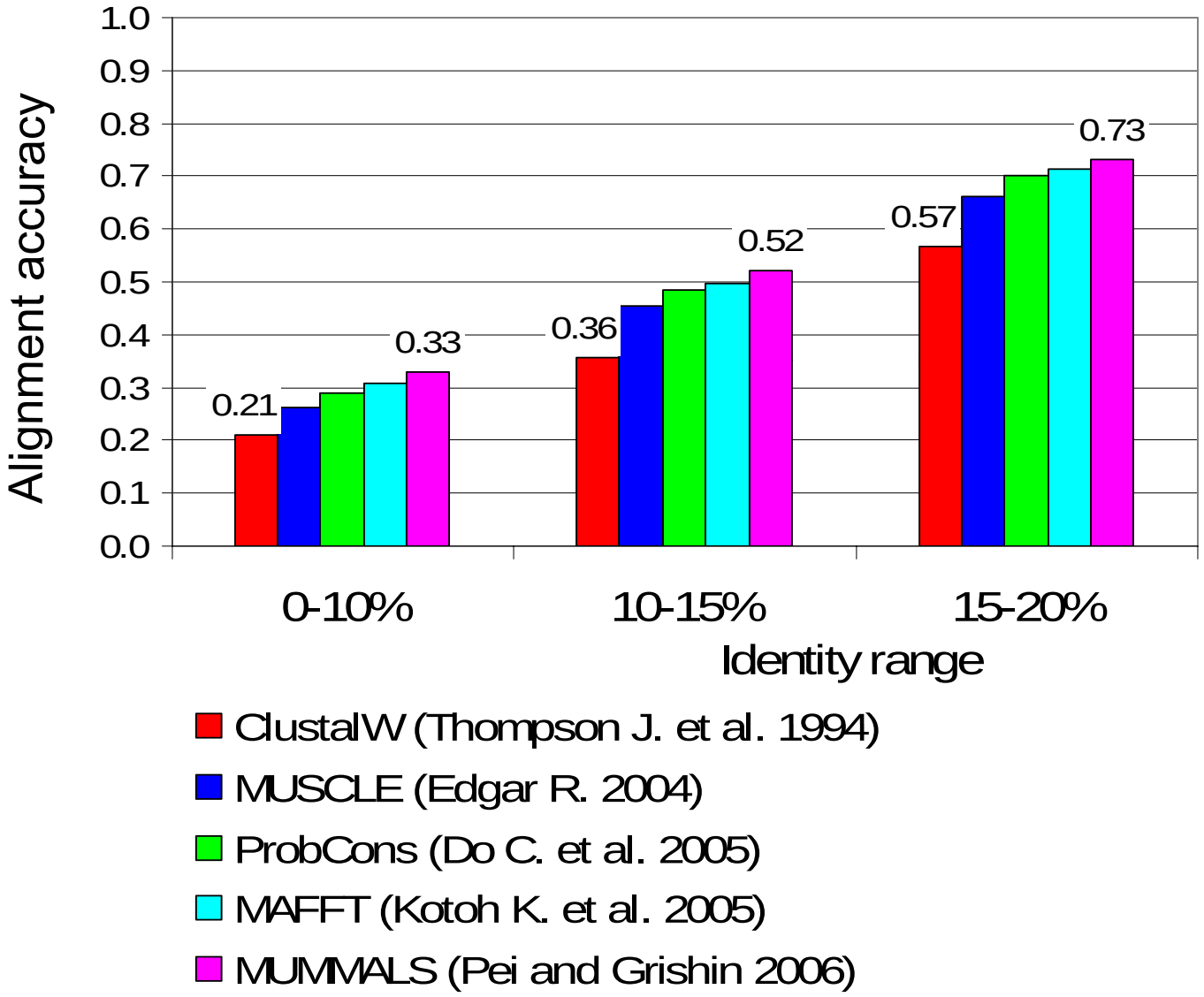
DALI superposition

ClustalW alignment accuracy



Tests on 1785 domain pairs from SCOP (Murzin A. et al. 1995) database.

What about other methods?

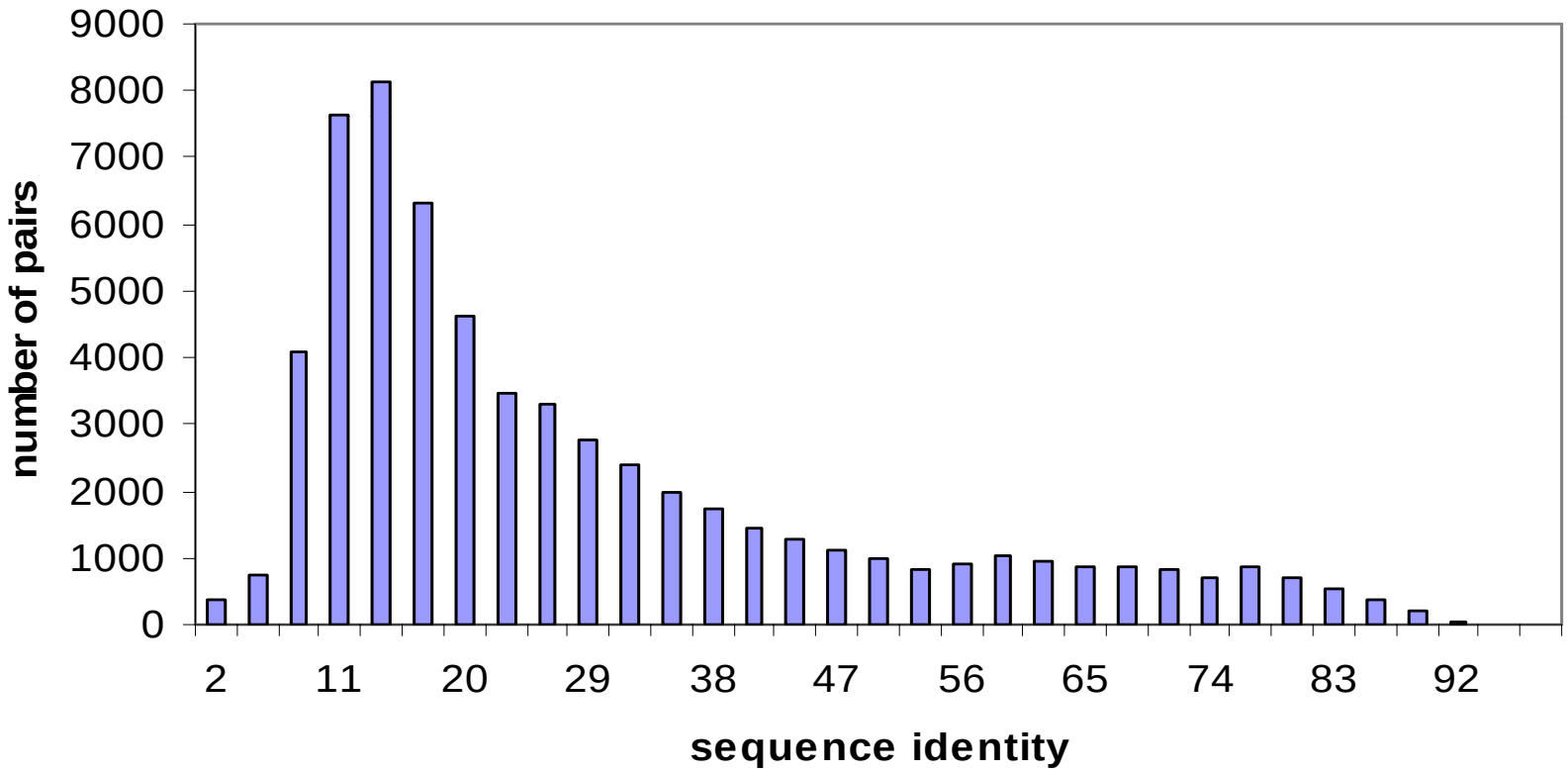


Why do we care about remote homologs,

i.e. alignments of sequence pairs with identity less than 20% ?

Why do we care about remote homologs? **Reason 1**

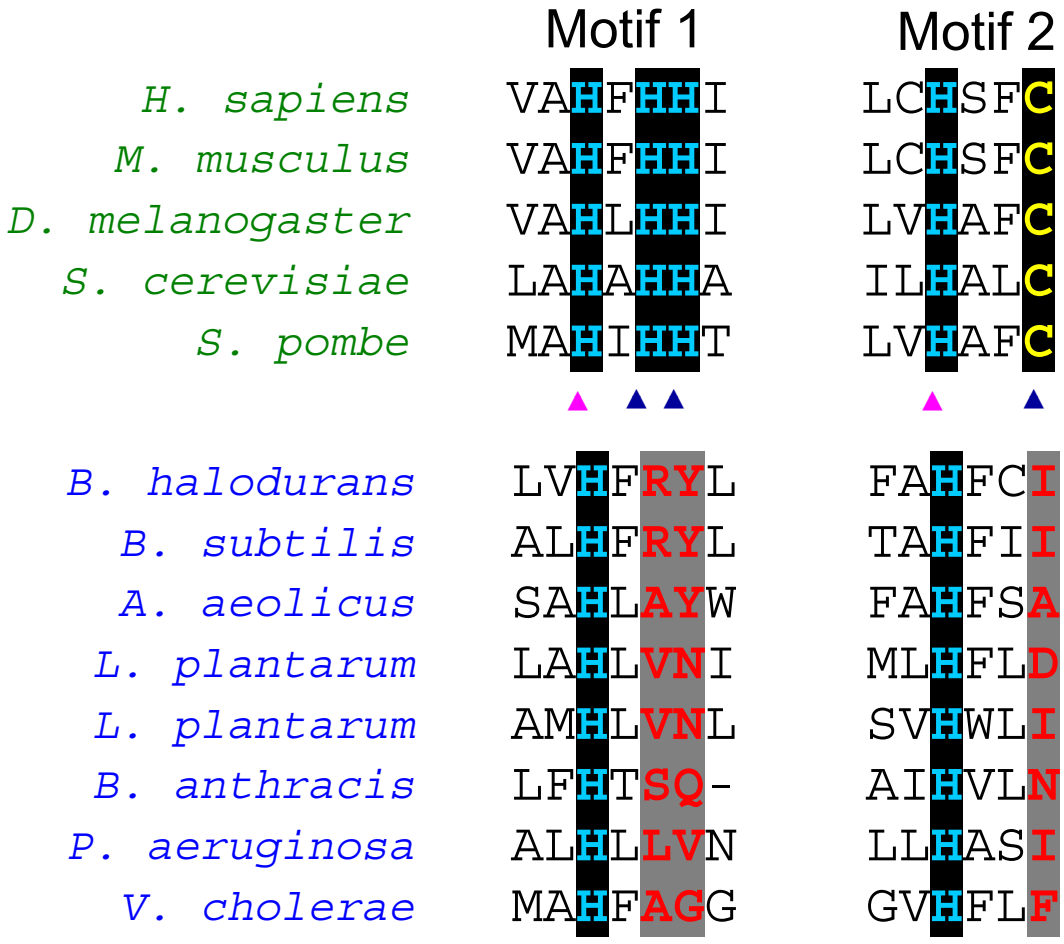
Sequence identity distribution for proteins with significant structural similarity (Dali Z-score >7.0) in FSSP[1] database



1. Holm and Sander, 1996

Why do we care about remote homologs? **Reason 2**

Distant homologs help prediction of functional residues



- RCEs are **CAAX prenyl proteases** identified in **eukaryotes**. (Dolence *et al*. 2000)

- Computational methods identified distant homologs of RCEs in many **bacteria**. (Pei and Grishin, 2001)

- Recent mutagenesis studies confirmed our predictions. (Plummer *et al*. 2005)

▲ mutations result in complete loss of activity

▲ mutations do not affect enzyme activity

Our goal (for a few years) has been

to improve alignment quality of distantly related sequences

PROMALS – PROfile Multiple Alignment with predicted Local Structure

PROMALS input: unaligned protein sequences

PROMALS output: multiple sequence alignment

PROMALS algorithm builds alignment of distantly related sequences by utilizing three main sources:

1. Predicted secondary structure
2. Homologous sequences from database searches
3. Complex but reasonable probabilistic models

Source #1: secondary structure

Secondary structure is more conserved than sequence

DALI structural alignment colored by real secondary structures

```
1w25 NRRYMTGQLDSLVKRATLGGDPVSALLIDIDFFKKINDTFGHDIGDEVLREFALRLASNVRA-IDLP-CRYGGEEFVVIMPDT-
1wc4 ------------PEPR----LITILFSDIVGFTRMSNALQSQGVAELLNEYLGEMTRAVFENQGTVDKFVG-DAIMALYGAPE

1w25 ------ALADALRIAERIRMHVSG-SPFTVAHGREMLN------VTISIGVSATAGEGDT----------PEALLKRADEGVYQ
1wc4 EMSPSEQVRRAIATARQMLVALEKLNQGW-QERGLVGRNEVPPVRFRCGIHQGMAVVGLFGSQERSDFTAIGPSVNIAARLQEA

1w25 AKASGRNAVVGKAA--------------------------------
1wc4 TA---PNSIMVSAMVAQYVPDEEIIKREFLELKGIDEPVMTCVINPN
```

sequence identity = 12%

Secondary structure prediction is about 80% accurate

DALI structural alignment colored by PSIPRED[1] predicted secondary structures

```
1w25 NRRYMTGQLDSLVKRATLGGDPVSALLIDIDFFKKINDTFGHDIGDEVLREFALRLASNVRA-IDLP-CRYGGEEFVVIMPDT-
1wc4 ------------PEPR----LITILFSDIVGFTRMSNALQSQGVAELLNEYLGEMTRAVFENQGTVDKFVG-DAIMALYGAPE

1w25 ------ALADALRIAERIRMHVSG-SPFTVAHGREMLN------VTISIGVSATAGEGDT----------PEALLKRADEGVYQ
1wc4 EMSPSEQVRRAIATARQMLVALEKLNQGW-QERGLVGRNEVPPVRFRCGIHQGMAVVGLFGSQERSDFTAIGPSVNIAARLQEA

1w25 AKASGRNAVVGKAA--------------------------------
1wc4 TA---PNSIMVSAMVAQYVPDEEIIKREFLELKGIDEPVMTCVINPN
```

1. Jones 1999

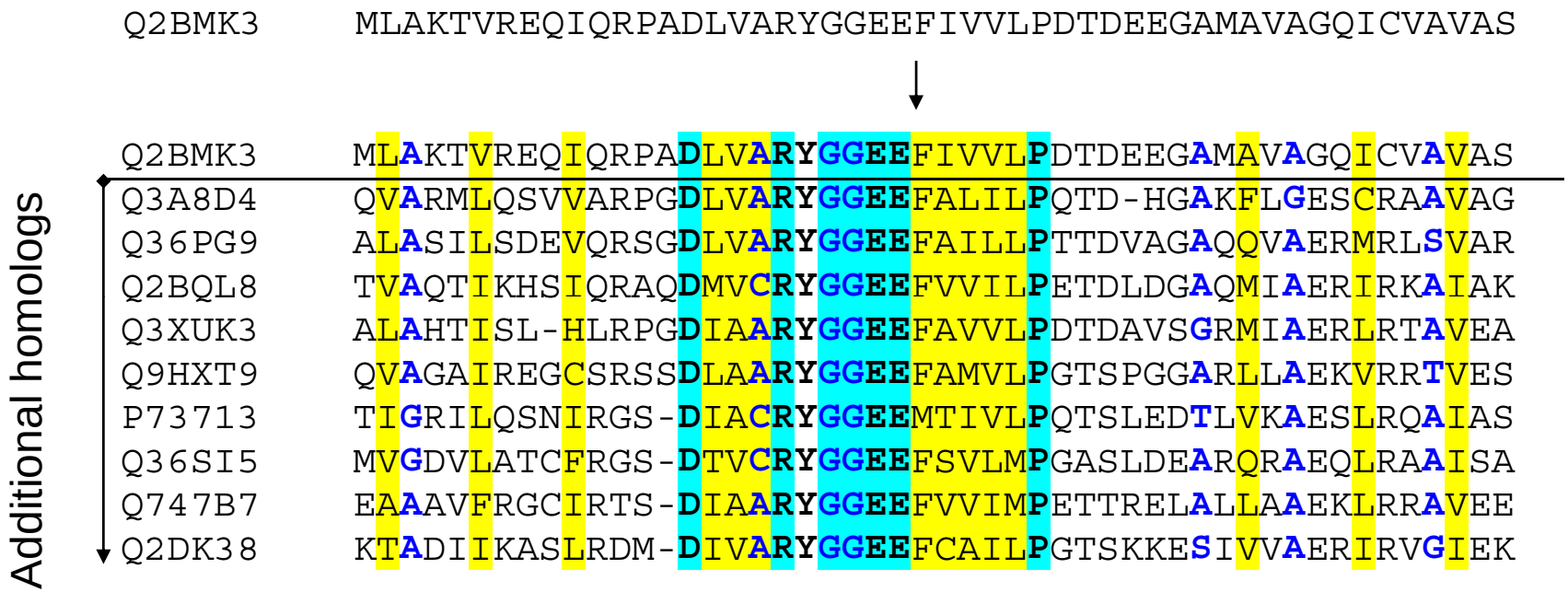
Source #2: homologous sequences

More homologs bring up important sequence features through averaging

```
Q2BMK3     MLAKTVREQIQRPADLVARYGGEEFIVVLPDTDEEGAMAVAGQICVAVAS
```

Source #2: homologous sequences

More homologs bring up important sequence features through averaging



A profile derived from multiple sequence alignment contains position-specific information about:

(1) amino acid usage

(2) amino acid conservation

Cyan: invariant position

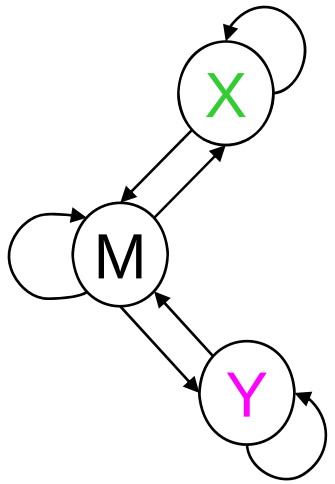
Yellow: hydrophobic position

Blue: small residues

Source #3: logical statistical models

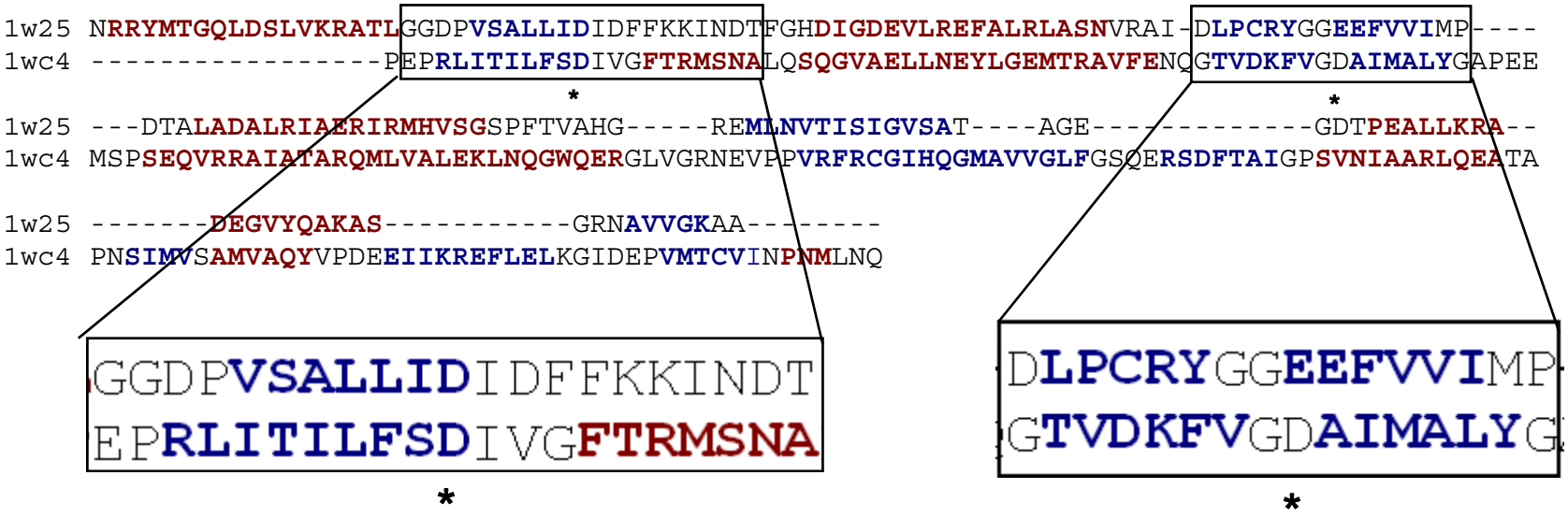
Statistical models of profile-profile alignment

Hidden Markov model



M: emit an aligned position pair
X: emit a position in first profile
Y: emit a position in second profile

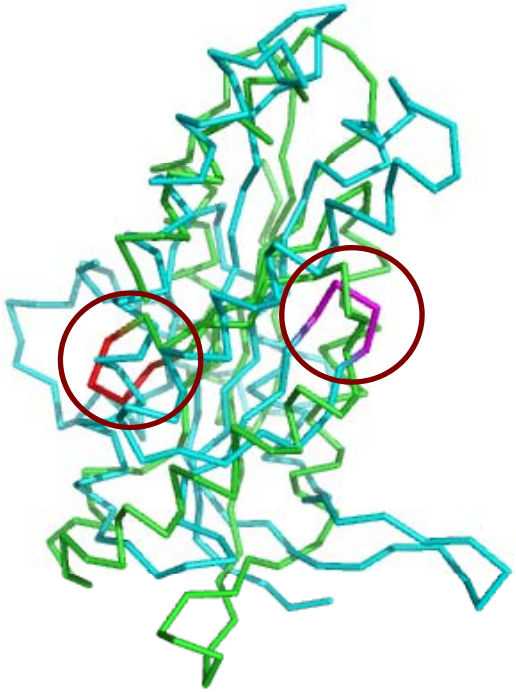
PROMALS alignment example: diguanylate cyclase and adenylate cyclase



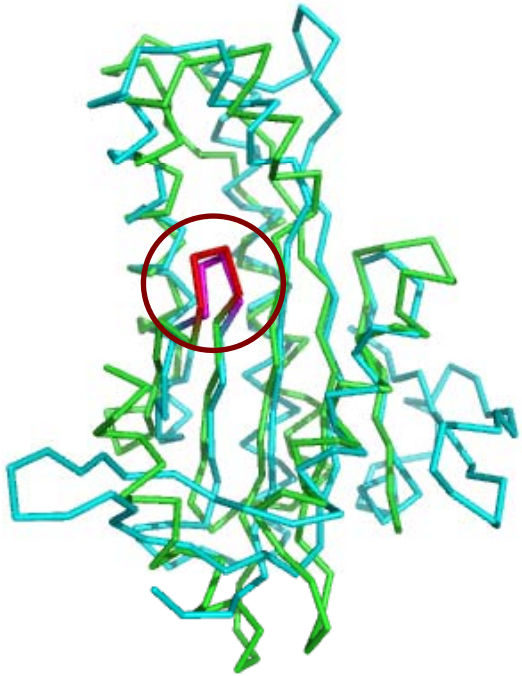
Red: predicted alpha-helix

Blue: predicted beta-strand

*: metal-binding residues



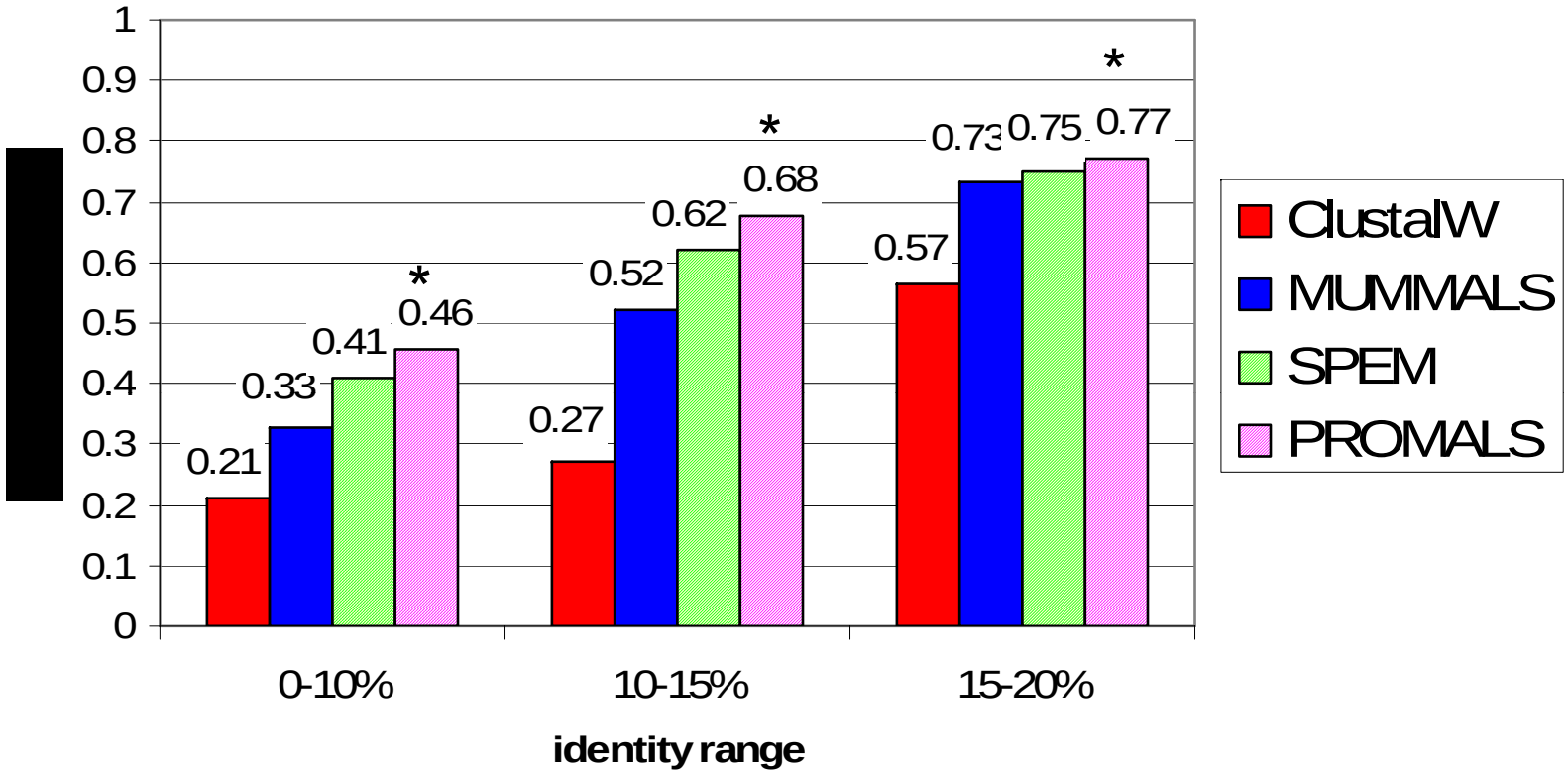
ClustalW superposition



DALI structural superposition

PROMALS superposition

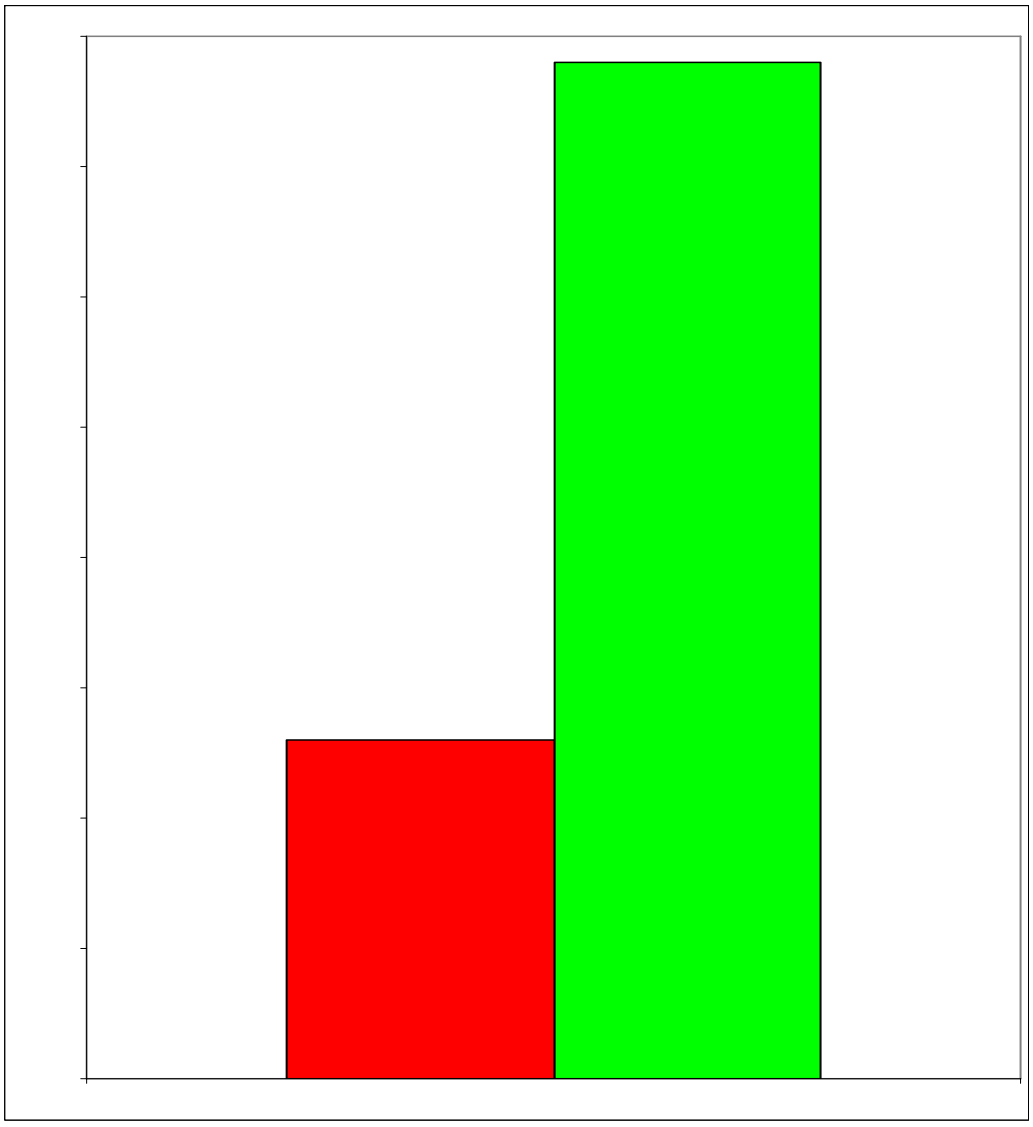
Tests on SCOP domain pairs binned by sequence identity



ClustalW and MUMMALS: methods that do not use additional homologs and predicted secondary structures.
SPEM and PROMALS: methods that use additional homologs and predicted secondary structures.

* PROMALS is statistically better than other methods (P<0.0001)

How accurate are PROMALS alignments?

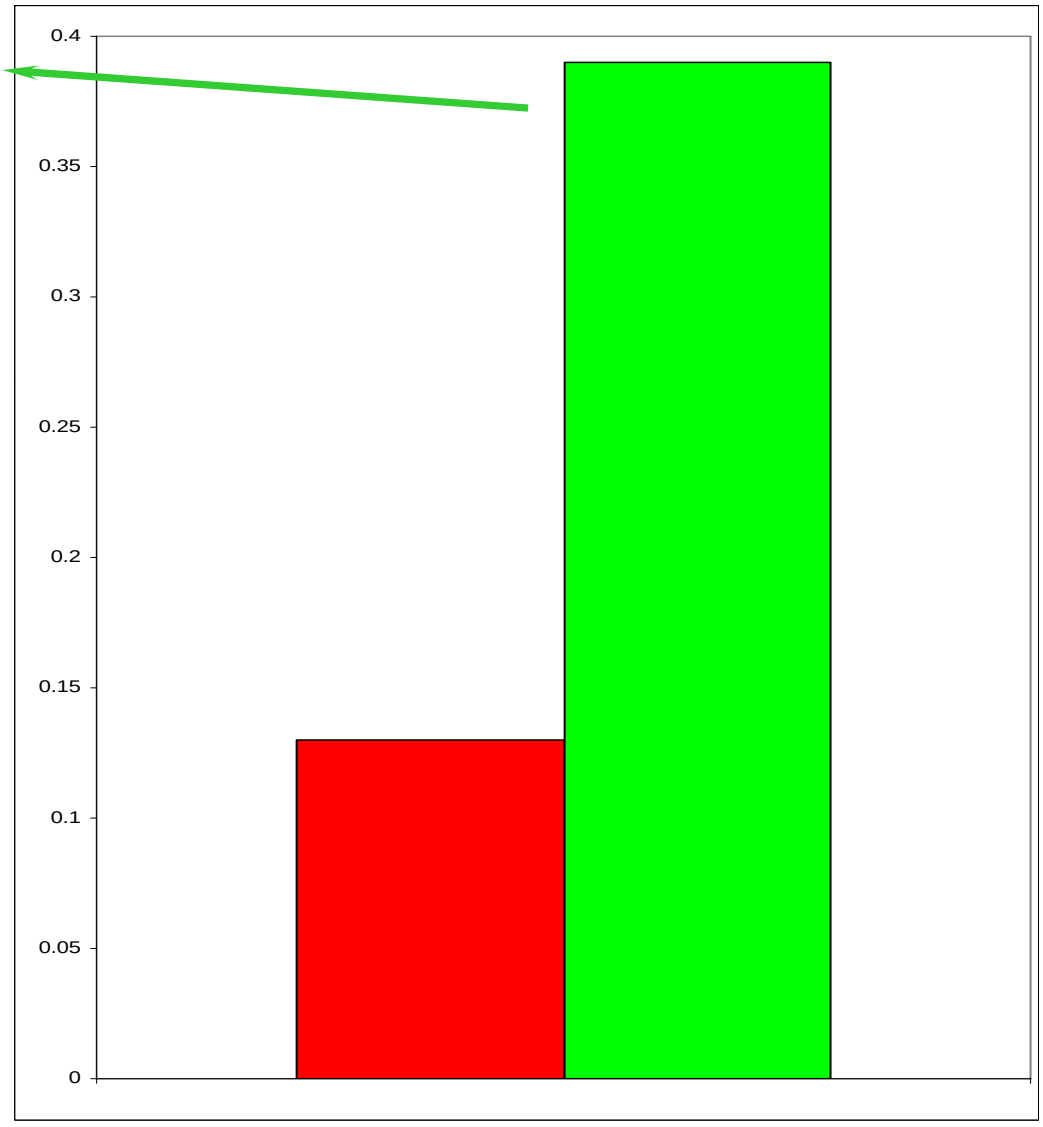


How accurate are PROMALS alignments?

40%

Accuracy for sequence pairs with

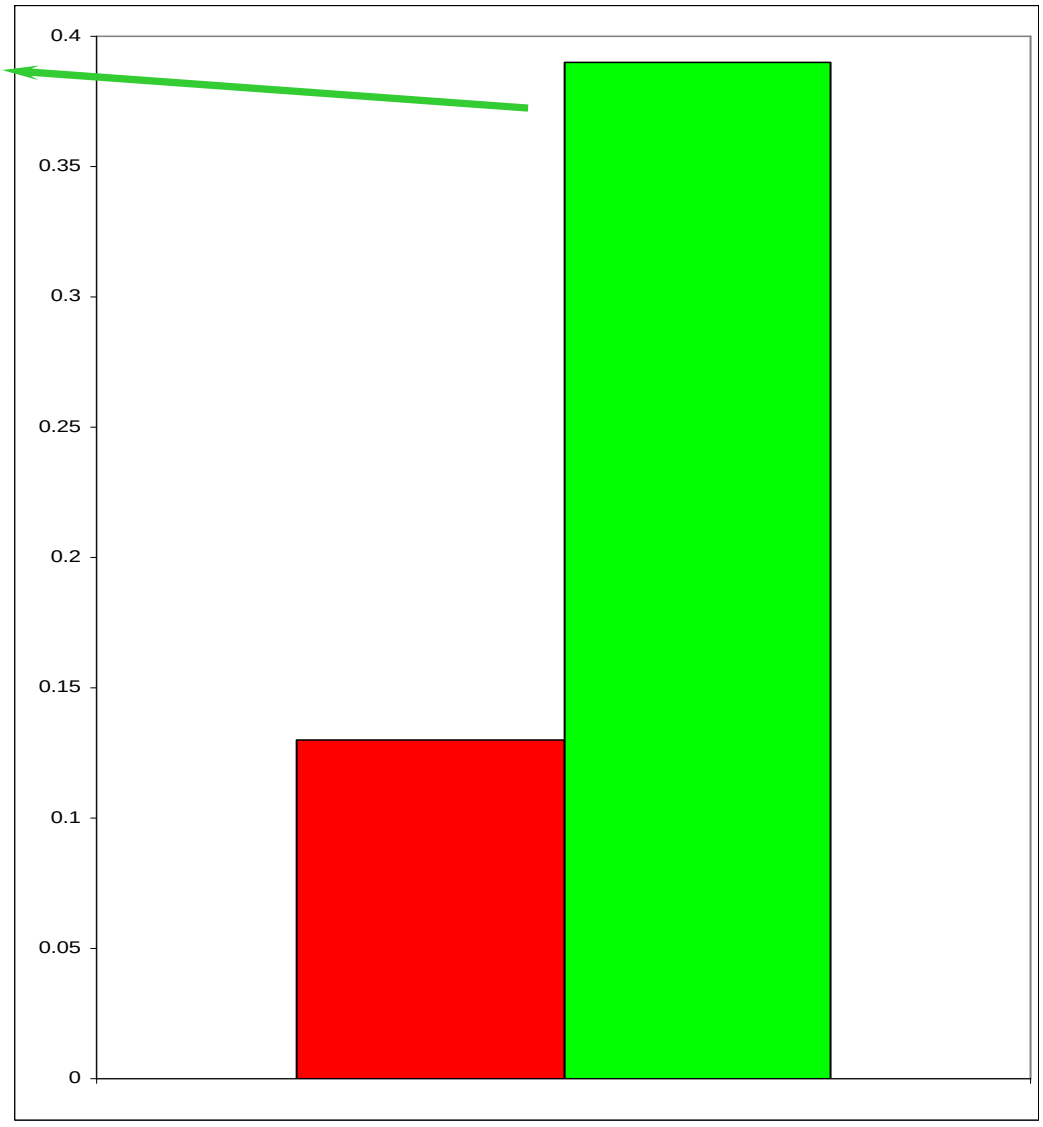
~7% identity



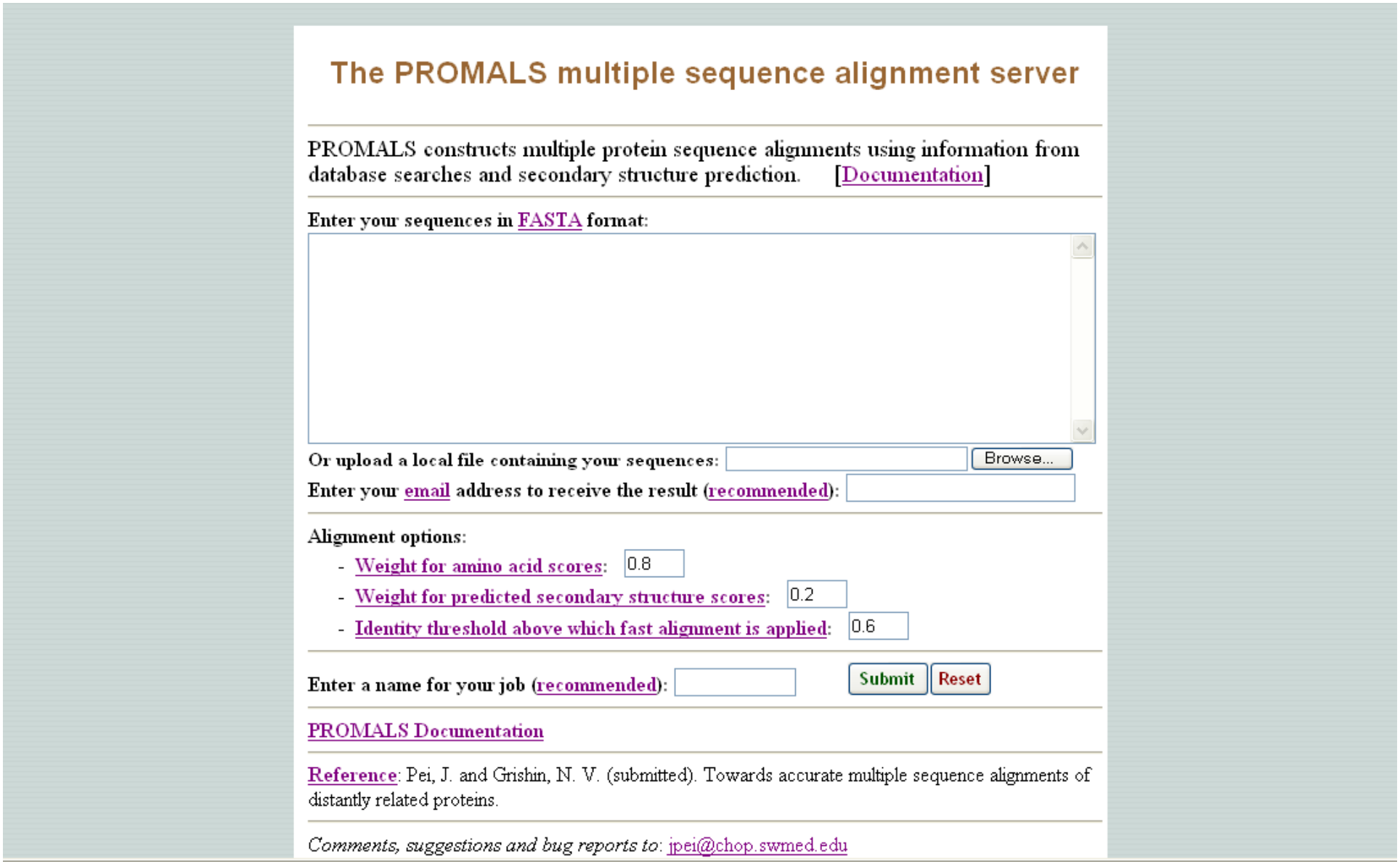
How accurate are PROMALS alignments?

40%

Accuracy for sequence pairs with
~7% identity



http://prodata.swmed.edu/promals



The PROMALS multiple sequence alignment server

PROMALS constructs multiple protein sequence alignments using information from database searches and secondary structure prediction. [Documentation]

Enter your sequences in FASTA format:

Or upload a local file containing your sequences: Browse...

Enter your email address to receive the result (recommended):

Alignment options:

- **Weight for amino acid scores:** 0.8
- **Weight for predicted secondary structure scores:** 0.2
- **Identity threshold above which fast alignment is applied:** 0.6

Enter a name for your job (recommended): Submit Reset

PROMALS Documentation

Reference: Pei, J. and Grishin, N. V. (submitted). Towards accurate multiple sequence alignments of distantly related proteins.

Comments, suggestions and bug reports to: jpei@chop.swmed.edu

What PROMALS does not do

It does not use

explicit 3D structure modeling techniques

It uses only input sequences and internal sequence database

It predicts secondary structure from sequences, but does not build 3D models

We have a decent alignment program,

where is the catch?

SPEED (or the lack of it) !

ClustalW takes seconds to minutes per alignment

PROMALS takes minutes to hours per alignment:

average is about **30 min per family**,
some large families take much, much longer

We have a decent alignment program, what NOT to do with it?

GI-GO effects:

non-homologous proteins should not be an input

Low complexity proteins should not be an input:

NQQQQQNNNSSSQQQQQQQQQQSSTTTTQQQQQQQQQNN

since the concept of an alignable position that can be traced to a common ancestor does not apply to them

Membrane proteins should be used with caution, since their amino acid composition is different, and we still have too few structures of them to test our algorithms thoroughly

Acknowledgement

Our group

Lisa Kinch
Jimin Pei
Sara Cheek
Shuoyong Shi
Indraneel M.
Yong Wang
Yi Zhong
Wei Cai
Erik Nelson
Ming Tang
Yuan Qi
Jamie Wrabl
Ruslan Sadreyev
Hua Cheng
Bong-Hyun Kim
Dorothee Staber

Collaborators

Eugene Koonin NCBI, NIH
Yuri Wolf NCBI, NIH
Eugene Shakhnovich Harvard
Andrei Osterman Burnham
Leszek Rychlewski Bioinfobank,Poland

HHMI, NIH, UTSW,
The Welch Foundation