Protein sequence analysis

Fasta format:

>1-line

AA sequence (1 or more lines)

```
>gi|22547186|ref|NP_004160.3| serine hydroxymethyltransferase,
cytosolic isoform 1 [Homo sapiens]
MTMPVNGAHKDADLWSSHDKMLAQPLKDSDVEVYNIIKKESNRQRVGLELIASENFASRAVLEALGSCLN
NKYSEGYPGQRYYGGTEFIDELETLCQKRALQAYKLDPQCWGVNVQPYSGSPANFAVYTALVEPHGRIMG
LDLPDGGHLTHGFMTDKKKISATSIFFESMPYKVNPDTGYINYDQLEENARLFHPKLIIAGTSCYSRNLE
YARLRKIADENGAYLMADMAHISGLVAAGVVPSPFEHCHVVTTTTHKTLRGCRAGMIFYRKGVKSVDPKT
GKEILYNLESLINSAVFPGLQGGPHNHAIAGVAVALKQAMTLEFKVYQHQVVANCRALSEALTELGYKIV
TGGSDNHLILVDLRSKGTDGGRAEKVLEACSIACNKNTCPGDRSALRPSGLRLGTPALTSRGLLEKDFQK
VAHFIHRGIELTLQIQSDTGVRATLKEFKERLAGDKYQAAVQALREEVESFASLFPLPGLPDF
```

Protein sequence analysis

Most proteins are modular

Domains: structural, functional, folding and evolutionary units

(30-700 a.a.; 100 a.a. on average)

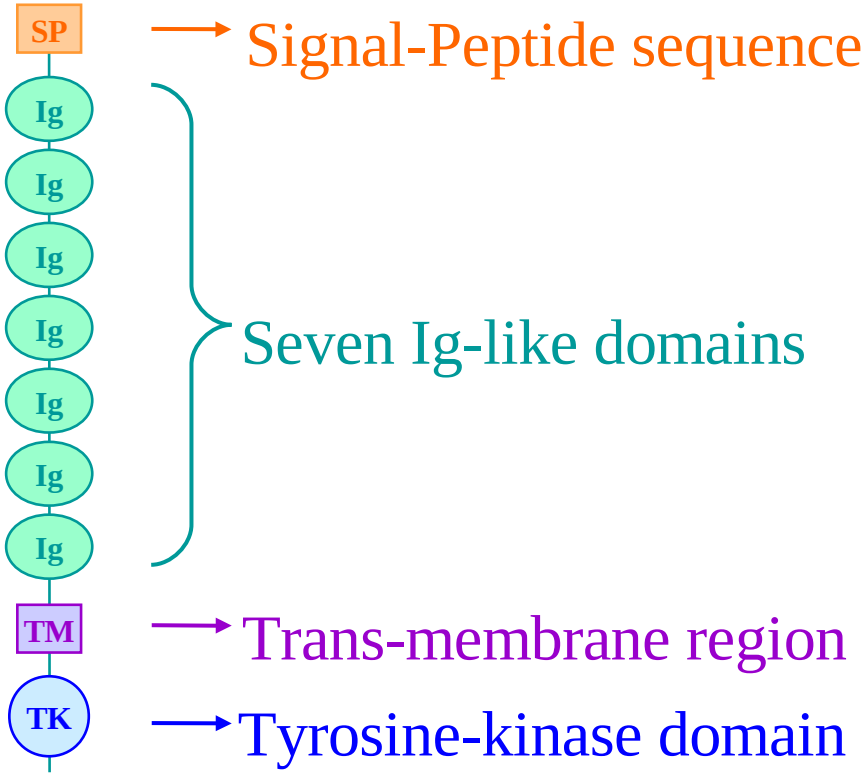
Analysis and prediction: domain – not whole protein – level

Protein sequence analysis

Proteins are modular

Domains: structural, functional, folding and evolutionary units

```
>gi|156104876|ref|NP_002010.2| vascular endothelial
growth factor receptor 1 isoform 1 precursor [Homo
sapiens] 5T89
MVSYWDTGVLLCALLSCLLLTGSSSGSKLKDPELSLKGTQHIMQAGQTLHLQCRG
EAAHKWSLPEMVSKESERLSITKSACGRNGKQFCSTLTLNTAQANHTGFYSCKYL
AVPTSKKKETESAIYIFISDTGRPFVEMYSEIPEIIHMTEGRELVIPCRVTSPNI
TVTLKKFPLDTLIPDGKRIIWDSRKGFIISNATYKEIGLLTCEATVNGHLYKTNY
LTHRQTNTIIDVQISTPRPVKLLRGHTLVLNCTATTPLNTRVQMTWSYPDEKNKR
ASVRRRIDQSNSHANIFYSVLTIDKMQNKDKGLYTCRVRSGPSFKSVNTSVHIYD
KAFITVKHRKQQVLETVAGKRSYRLSMKVKAFPSPEVVWLKDGLPATEKSARYLT
RGYSLIIKDVTEEDAGNYTILLSIKQSNVFKNLTATLIVNVKPQIYEKAVSSFPD
PALYPLGSRQILTCTAYGIPQPTIKWFWHPCNHNHSEARCDFCSNNEESFILDAD
SNMGNRIESITQRMAIIEGKNKMASTLVVADSRISGIYICIASNKVGTVGRNISF
YITDVPNGFHVNLEKMPTEGEDLKLSCTVNKFLYRDVTWILLRTVNNRTMHYSIS
KQKMAITKEHSITLNLTIMNVSLQDSGTYACRARNVYTGEEILQKKEITIRDQEA
PYLLRNLSDHTVAISSSTTLDCHANGVPEPQITWFKNNHKIQQEPGIILGPGSST
LFIERVTEEDEGVYHCKATNQKGSVESSAYLTVQGTSDKSNLELITLTCTCVAAT
LFWLLLTLFIRKMKRSSSEIKTDYLSIIMDPDEVPLDEQCERLPYDASKWEFARE
RLKLGKSLGRGAFGKVVQASAFGIKKSPTCRTVAVKMLKEGATASEYKALMTELK
ILTHIGHHLNVVNLLGACTKQGGPLMVIVEYCKYGNLSNYLKSKRDLFFLNKDAA
LHMEPKKEKMEPGLEQGKKPRLDSVTSSESFASSGFQEDKSLSDVEEEEDSDGFY
KEPITMEDLISYSFQVARGMEFLSSRKCIHRDLAARNILLSENNVVKICDFGLAR
DIYKNPDYVRKGDTRLPLKWMAPESIFDKIYSTKSDVWSYGVLLWEIFSLGGSPY
PGVQMDEDFCSRLREGMRMRAPEYSTPEIYQIMLDCWHRDPKERPRFAELVEKLG
DLLQANVQQDGKDYIPINAILTGNSGFTYSTPAFSEDFFKESISAPKFNSGSSDD
VRYVNAFKFMSLERIKTFEELLPNATSMFDDYQGDSSTLLASPMLKRFTWTDSKP
KASLKIDLRVTSKSKESGLSDVSRPSFCHSSCGHVSEGKRRFTYDHAELERKIAC
CSPPPDYNSVVLYSTPPI
```



Protein sequence analysis

1) Save Fasta sequence
2) Run BLAST
    - Parameters:
        - Max target sequences (5000)
        - Organism
        - Expect threshold
        - Filter low-complexity regions

Protein sequence analysis

Program output

1.) Conserved domains (CDD)
& Active/binding sites

Protein sequence analysis

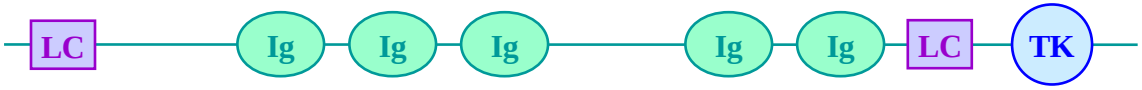
Identify protein domains

- **3D structure** (Blast vs. PDB)
- **CDD** (NCBI)
- **Pfam**: pfam.sanger.ac.uk
- **SMART**: smart.embl-heidelberg.de
- **Superfamily**: supfam.cs.bris.ac.uk/SUPERFAMILY/
- …

Protein sequence analysis

Domain prediction: CDD

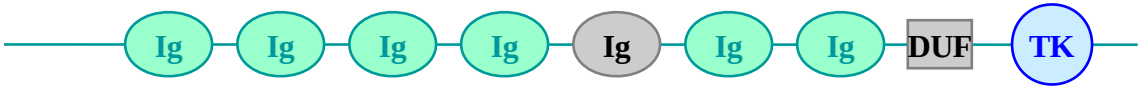
- 5 Immunoglobulin (Ig)-like domains
- Protein tyrosine kinase catalytic domain



Protein sequence analysis

Domain prediction: Pfam

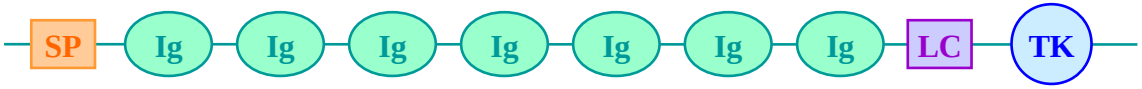
- 6 Ig-like domains (+1 below threshold)
- Protein tyrosine kinase (TK) catalytic domain
- Domain of unknown function (below threshold)



Protein sequence analysis

Domain prediction: SMART

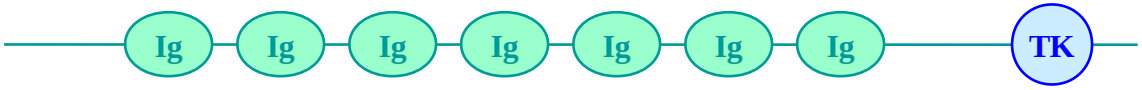
- 7 Ig-like domains
- Protein tyrosine kinase (TK) catalytic domain
- Signal peptide (SP)
- Low complexity region (LC)



Protein sequence analysis

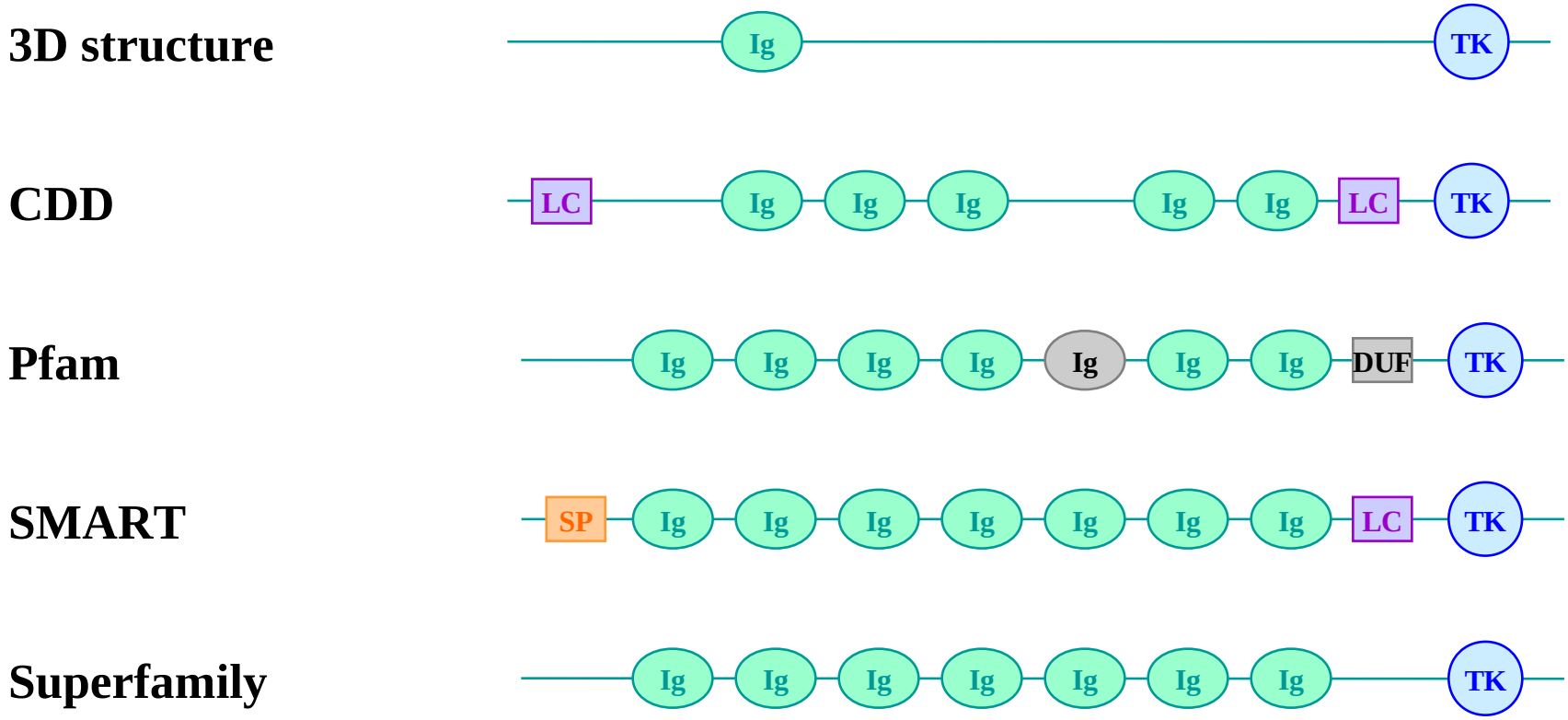
Domain prediction: Superfamily

- 7 Ig-like domains
- Protein tyrosine kinase (TK) catalytic domain



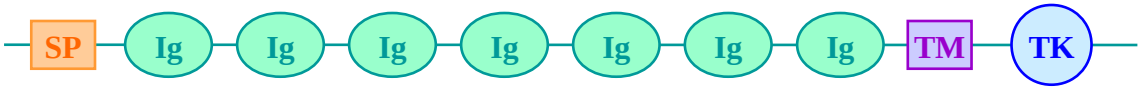
Protein sequence analysis

Results summary



Protein sequence analysis

Map predicted domains on your sequence

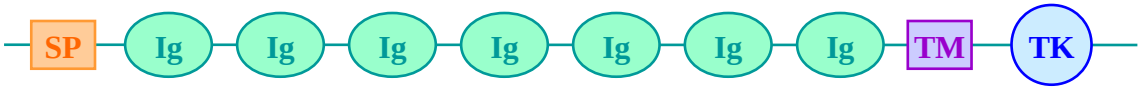


>gi|156104876|ref|NP_002010.2| vascular endothelial growth factor receptor 1 isoform 1 precursor [Homo sapiens]

MVSYWDTGVL LCALLSCLLL TGSSSGSKLK DPELSLKGTQ HIMQAGQTLH LQCRGEAAHK
WSLPEMVSKE SERLSITKSA CGRNGKQFCS TLTLNTAQAN HTGFYSCKYL AVPTSKKKET
ESAIYIFISD TGRPFVEMYS EIPEIIHMTE GRELVIPCRV TSPNITVTLK KFPLDTLIPD
GKRIIWDSRK GFIISNATYK EIGLLTCEAT VNGHLYKTNY LTHRQTNTII DVQISTPRPV
KLLRGHTLVL NCTATTPLNT RVQMTWSYPD EKNKRASVRR RIDQSNSHAN IFYSVLTIDK
MQNKDKGLYT CRVRSGPSFK SVNTSVHIYD KAFITVKHRK QQVLETVAGK RSYRLSMKVK
AFPSPEVVWL KDGLPATEKS ARYLTRGYSL IIKDVTEEDA GNYTILLSIK QSNVFKNLTA
TLIVNVKPQI YEKAVSSFPD PALYPLGSRQ ILTCTAYGIP QPTIKWFWHP CNHNHSEARC
DFCSNNEESF ILDADSNMGN RIESITQRMA IIEGKNKMAS TLVVADSRIS GIYICIASNK
VGTVGRNISF YITDVPNGFH VNLEKMPTEG EDLKLSCTVN KFLYRDVTWI LLRTVNNRTM
HYSISKQKMA ITKEHSITLN LTIMNVSLQD SGTYACRARN VYTGEEILQK KEITIRDQEA
PYLLRNLSDH TVAISSSTTL DCHANGVPEP QITWFKNNHK IQQEPGIILG PGSSTLFIER
VTEEDEGVYH CKATNQKGSV ESSAYLTVQG TSDKSNLELI TLTCTCVAAT LFWLLLTLFI
RKMKRSSSEI KTDYLSIIMD PDEVPLDEQC ERLPYDASKW EFARERLKLG KSLGRGAFGK
VVQASAFGIK KSPTCRTVAV KMLKEGATAS EYKALMTELK ILTHIGHHLN VVNLLGACTK
QGGPLMVIVE YCKYGNLSNY LKSKRDLFFL NKDAALHMEP KKEKMEPGLE QGKKPRLDSV
TSSESFASSG FQEDKSLSDV EEEEDSDGFY KEPITMEDLI SYSFQVARGM EFLSSRKCIH
RDLAARNILL SENNVVKICD FGLARDIYKN PDYVRKGDTR LPLKWMAPES IFDKIYSTKS
DVWSYGVLLW EIFSLGGSPY PGVQMDEDFC SRLREGMRMR APEYSTPEIY QIMLDCWHRD
PKERPRFAEL VEKLGDLLQA NVQQDGKDYI PINAILTGNS GFTYSTPAFS EDFFKESISA
PKFNSGSSDD VRYVNAFKFM SLERIKTFEE LLPNATSMFD DYQGDSSTLL ASPMLKRFTW
TDSKPKASLK IDLRVTSKSK ESGLSDVSRP SFCHSSCGHV SEGKRRFTYD HAELERKIAC
CSPPPDYNSV VLYSTPPI

Protein sequence analysis

Divide your sequence into potential domain fragments

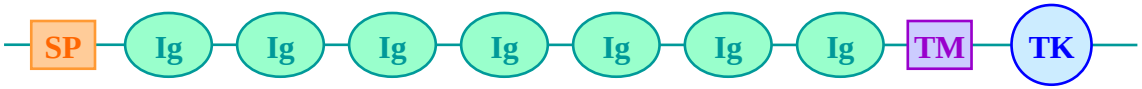


>gi|156104876|ref|NP_002010.2| VEGFR-1 [Homo sapiens]
MVSYWDTGVLLCALLSCLLLTGSSSGSKLKDPELSLKGTQHIMQAGQTLHLQCRGEAAHKWSLPEMVSKESERLSITKSACGRNGKQFCSTLTLNTAQANHTGFYSCKYLAVPTSKKKETESAIYIFISDTGRPFVEM
YSEIPEIIHMTEGRELVIPCRVTSPNITVTLKKFPLDTLIPDGKRIIWDSRKGFIISNATYKEIGLLTCEATVNGHLYKTNYLTHRQTNTIIDVQISTPRPVKLLRGHTLVLNCTATTPLNTRVQMTWSYPDEKNKRA
SVRRRIDQSNSHANIFYSVLTIDKMQNKDKGLYTCRVRSGPSFKSVNTSVHIYDKAFITVKHRKQQVLETVAGKRSYRLSMKVKAFPSPEVVWLKDGLPATEKSARYLTRGYSLIIKDVTEEDAGNYTILLSIKQSNV
FKNLTATLIVNVKPQIYEKAVSSFPDPALYPLGSRQILTCTAYGIPQPTIKWFWHPCNHNHSEARCDFCSNNEESFILDADSNMGNRIESITQRMAIIEGKNKMASTLVVADSRISGIYICIASNKVGTVGRNISFYI
TDVPNGFHVNLEKMPTEGEDLKLSCTVNKFLYRDVTWILLRTVNNRTMHYSISKQKMAITKEHSITLNLTIMNVSLQDSGTYACRARNVYTGEEILQKKEITIRDQEAPYLLRNLSDHTVAISSSTTLDCHANGVPEP
QITWFKNNHKIQQEPGIILGPGSSTLFIERVTEEDEGVYHCKATNQKGSVESSAYLTVQGTSDKSNLELITLTCTCVAATLFWLLLTLFIRKMKRSSSEIKTDYLSIIMDPDEVPLDEQCERLPYDASKWEFARERLK
LGKSLGRGAFGKVVQASAFGIKKSPTCRTVAVKMLKEGATASEYKALMTELKILTHIGHHLNVVNLLGACTKQGGPLMVIVEYCKYGNLSNYLKSKRDLFFLNKDAALHMEPKKEKMEPGLEQGKKPRLDSVTSSESF
ASSGFQEDKSLSDVEEEEDSDGFYKEPITMEDLISYSFQVARGMEFLSSRKCIHRDLAARNILLSENNVVKICDFGLARDIYKNPDYVRKGDTRLPLKWMAPESIFDKIYSTKSDVWSYGVLLWEIFSLGGSPYPGVQ
MDEDFCSRLREGMRMRAPEYSTPEIYQIMLDCWHRDPKERPRFAELVEKLGDLLQANVQQDGKDYIPINAILTGNSGFTYSTPAFSEDFFKESISAPKFNSGSSDDVRYVNAFKFMSLERIKTFEELLPNATSMFDDY
QGDSSTLLASPMLKRFTWTDSKPKASLKIDLRVTSKSKESGLSDVSRPSFCHSSCGHVSEGKRRFTYDHAELERKIACCSPPPDYNSVVLYSTPPI

>Ig-like-1	>Ig-like-2	>Ig-like-3	>Ig-like-4	>Ig-like-5	>Ig-like-6	>Ig-like-7	>TK
LLLTGSSSGSK	AIYIFISDTGR	KTNYLTHRQTN	TSVHIYDKAFI	ATLIVNVKPQI	ISFYITDVPNG	ITIRDQEAPYL	DEQCERLPYDASKWEFARERLKLGK
LKDPELSLKGT	PFVEMYSEIPE	TIIDVQISTPR	TVKHRKQQVLE	YEKAVSSFPDP	FHVNLEKMPTE	LRNLSDHTVAI	SLGRGAFGKVVQASAFGIKKSPTCR
QHIMQAGQTLH	IIHMTEGRELV	PVKLLRGHTLV	TVAGKRSYRLS	ALYPLGSRQIL	GEDLKLSCTVN	SSSTTLDCHAN	TVAVKMLKEGATASEYKALMTELKI
LQCRGEAAHKW	IPCRVTSPNIT	LNCTATTPLNT	MKVKAFPSPEV	TCTAYGIPQPT	KFLYRDVTWIL	GVPEPQITWFK	LTHIGHHLNVVNLLGACTKQGGPLM
SLPEMVSKESE	VTLKKFPLDTL	RVQMTWSYPDE	VWLKDGLPATE	IKWFWHPCNHN	LRTVNNRTMHY	NNHKIQQEPGI	VIVEYCKYGNLSNYLKSKRDLFFLN
RLSITKSACGR	IPDGKRIIWDS	KNKRASVRRRI	KSARYLTRGYS	HSEARCDFCSN	SISKQKMAITK	ILGPGSSTLFI	KDAALHMEPKKEKMEPGLEQGKKPR
NGKQFCSTLTL	RKGFIISNATY	DQSNSHANIFY	LIIKDVTEEDA	NEESFILDADS	EHSITLNLTIM	ERVTEEDEGVY	LDSVTSSESFASSGFQEDKSLSDVE
NTAQANHTGFY	KEIGLLTCEAT	SVLTIDKMQNK	GNYTILLSIKQ	NMGNRIESITQ	NVSLQDSGTYA	HCKATNQKGSV	EEEDSDGFYKEPITMEDLISYSFQV
SCKYLAVPTSK	VNGHLYKTNYL	DKGLYTCRVRS	SNVFKNLTATL	RMAIIEGKNKM	CRARNVYTGEE	ESSAYLTVQGT	ARGMEFLSSRKCIHRDLAARNILLS
KKETESAIYIF	THRQTNTIIDV	GPSFKSVNTSV	IVNVKPQIYEK	ASTLVVADSRI	ILQKKEITIRD	SDKSNLE	ENNVVKICDFGLARDIYKNPDYVRK
ISDTGRPFVEM	QISTPRPVKLL	HIYDKAFITVK	AVSSFPDPALY	SGIYICIASNK	QEAPYL		GDTRLPLKWMAPESIFDKIYSTKSD
YSEIPEIIHMT	R	HRKQQVLETVA	P	VGTVGRNISFY			VWSYGVLLWEIFSLGGSPYPGVQMD
		G		ITDVPNGFHVN			EDFCSRLREGMRMRAPEYSTPEIYQ
				LEKMPTEGEDL			IMLDCWHRDPKERPRFAELVEKLGD
							LLQANVQQDGKDYIPINA

Protein sequence analysis

Analyse each fragment separately



>Ig-like-1
LLLTGSSSGSK
LKDPELSLKGT
QHIMQAGQTLH
LQCRGEAAHKW
SLPEMVSKESE
RLSITKSACGR
NGKQFCSTLTL
NTAQANHTGFY
SCKYLAVPTSK
KKETESAIYIF
ISDTGRPFVEM
YSEIPEIIHMT

>Ig-like-2
AIYIFISDTGR
PFVEMYSEIPE
IIHMTEGRELV
IPCRVTSPNIT
VTLKKFPLDTL
IPDGKRIIWDS
RKGFIISNATY
KEIGLLTCEAT
VNGHLYKTNYL
THRQTNTIIDV
QISTPRPVKLL
R

>Ig-like-3
KTNYLTHRQTN
TIIDVQISTPR
PVKLLRGHTLV
LNCTATTPLNT
RVQMTWSYPDE
KNKRASVRRRI
DQSNSHANIFY
SVLTIDKMQNK
DKGLYTCRVRS
GPSFKSVNTSV
HIYDKAFITVK
HRKQQVLETVA
G

>Ig-like-4
TSVHIYDKAFI
TVKHRKQQVLE
TVAGKRSYRLS
MKVKAFPSPEV
VWLKDGLPATE
KSARYLTRGYS
LIIKDVTEEDA
GNYTILLSIKQ
SNVFKNLTATL
IVNVKPQIYEK
AVSSFPDPALY
P

>Ig-like-5
ATLIVNVKPQI
YEKAVSSFPDP
ALYPLGSRQIL
TCTAYGIPQPT
IKWFWHPCNHN
HSEARCDFCSN
NEESFILDADS
NMGNRIESITQ
RMAIIEGKNKM
ASTLVVADSRI
SGIYICIASNK
VGTVGRNISFY
ITDVPNGFHVN
LEKMPTEGEDL

>Ig-like-6
ISFYITDVPNG
FHVNLEKMPTE
GEDLKLSCTVN
KFLYRDVTWIL
LRTVNNRTMHY
SISKQKMAITK
EHSITLNLTIM
NVSLQDSGTYA
CRARNVYTGEE
ILQKKEITIRD
QEAPYL

>Ig-like-7
ITIRDQEAPYL
LRNLSDHTVAI
SSSTTLDCHAN
GVPEPQITWFK
NNHKIQQEPGI
ILGPGSSTLFI
ERVTEEDEGVY
HCKATNQKGSV
ESSAYLTVQGT
SDKSNLE

>TK
DEQCERLPYDASKWEFARERLKLGK
SLGRGAFGKVVQASAFGIKKSPTCR
TVAVKMLKEGATASEYKALMTELKI
LTHIGHHLNVVNLLGACTKQGGPLM
VIVEYCKYGNLSNYLKSKRDLFFLN
KDAALHMEPKKEKMEPGLEQGKKPR
LDSVTSSESFASSGFQEDKSLSDVE
EEEDSDGFYKEPITMEDLISYSFQV
ARGMEFLSSRKCIHRDLAARNILLS
ENNVVKICDFGLARDIYKNPDYVRK
GDTRLPLKWMAPESIFDKIYSTKSD
VWSYGVLLWEIFSLGGSPYPGVQMD
EDFCSRLREGMRMRAPEYSTPEIYQ
IMLDCWHRDPKERPRFAELVEKLGD
LLQANVQQDGKDYIPINA

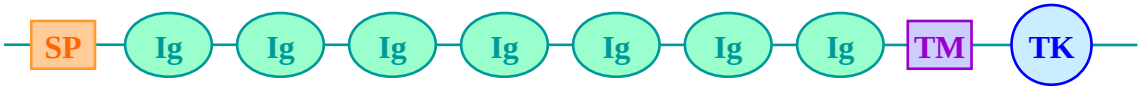
Analyse inter-domain sequences (domain: ~ 30 a.a.)

>Inter-domain-region
ITLTCTCVAATLFWLLLTLFIRKMKRSSSEIKTDYLSIIM
DPDEVPLDEQCERLPYDASKWEFARERLKLGKSLGRGAFG
KVVQASA

>C-ter-region
DCWHRDPKERPRFAELVEKLGDLLQANVQQDGKDYIPINAILTGNSGFTYST
PAFSEDFFKESISAPKFNSGSSDDVRYVNAFKFMSLERIKTFEELLPNATSM
FDDYQGDSSTLLASPMLKRFTWTDSKPKASLKIDLRVTSKSKESGLSDVSRP
SFCHSSCGHVSEGKRRFTYDHAELERKIACCSPPPDYNSVVLYSTPPI

Protein sequence analysis

Analyse each fragment separately



>Ig-like-1
LLLTGSSSGSK
LKDPELSLKGT
QHIMQAGQTLH
LQCRGEAAHKW
SLPEMVSKESE
RLSITKSACGR
NGKQFCSTLTL
NTAQANHTGFY
SCKYLAVPTSK
KKETESAIYIF
ISDTGRPFVEM
YSEIPEIIHMT

>Ig-like-2
AIYIFISDTGR
PFVEMYSEIPE
IIHMTEGRELV
IPCRVTSPNIT
VTLKKFPLDTL
IPDGKRIIWDS
RKGFIISNATY
KEIGLLTCEAT
VNGHLYKTNYL
THRQTNTIIDV
QISTPRPVKLL
R

>Ig-like-3
KTNYLTHRQTN
TIIDVQISTPR
PVKLLRGHTLV
LNCTATTPLNT
RVQMTWSYPDE
KNKRASVRRRI
DQSNSHANIFY
SVLTIDKMQNK
DKGLYTCRVRS
GPSFKSVNTSV
HIYDKAFITVK
HRKQQVLETVA
G

>Ig-like-4
TSVHIYDKAFI
TVKHRKQQVLE
TVAGKRSYRLS
MKVKAFPSPEV
VWLKDGLPATE
KSARYLTRGYS
LIIKDVTEEDA
GNYTILLSIKQ
SNVFKNLTATL
IVNVKPQIYEK
AVSSFPDPALY
P

>Ig-like-5
ATLIVNVKPQI
YEKAVSSFPDP
ALYPLGSRQIL
TCTAYGIPQPT
IKWFWHPCNHN
HSEARCDFCSN
NEESFILDADS
NMGNRIESITQ
RMAIIEGKNKM
ASTLVVADSRI
SGIYICIASNK
VGTVGRNISFY
ITDVPNGFHVN
LEKMPTEGEDL

>Ig-like-6
ISFYITDVPNG
FHVNLEKMPTE
GEDLKLSCTVN
KFLYRDVTWIL
LRTVNNRTMHY
SISKQKMAITK
EHSITLNLTIM
NVSLQDSGTYA
CRARNVYTGEE
ILQKKEITIRD
QEAPYL

>Ig-like-7
ITIRDQEAPYL
LRNLSDHTVAI
SSSTTLDCHAN
GVPEPQITWFK
NNHKIQQEPGI
ILGPGSSTLFI
ERVTEEDEGVY
HCKATNQKGSV
ESSAYLTVQGT
SDKSNLE

>TK
DEQCERLPYDASKWEFARERLKLGK
SLGRGAFGKVVQASAFGIKKSPTCR
TVAVKMLKEGATASEYKALMTELKI
LTHIGHHLNVVNLLGACTKQGGPLM
VIVEYCKYGNLSNYLKSKRDLFFLN
KDAALHMEPKKEKMEPGLEQGKKPR
LDSVTSSESFASSGFQEDKSLSDVE
EEEDSDGFYKEPITMEDLISYSFQV
ARGMEFLSSRKCIHRDLAARNILLS
ENNVVKICDFGLARDIYKNPDYVRK
GDTRLPLKWMAPESIFDKIYSTKSD
VWSYGVLLWEIFSLGGSPYPGVQMD
EDFCSRLREGMRMRAPEYSTPEIYQ
IMLDCWHRDPKERPRFAELVEKLGD
LLQANVQQDGKDYIPINA

Analyse inter-domain sequences (domain: ~ 30 a.a.)

>Inter-domain-region
ITLTCTCVAATLFWLLLTLFIRKMKRSSSEIKTDYLSIIM
DPDEVPLDEQCERLPYDASKWEFARERLKLGKSLGRGAFG
KVVQASA

>C-ter-region
DCWHRDPKERPRFAELVEKLGDLLQANVQQDGKDYIPINAILTGNSGFTYST
PAFSEDFFKESISAPKFNSGSSDDVRYVNAFKFMSLERIKTFEELLPNATSM
FDDYQGDSSTLLASPMLKRFTWTDSKPKASLKIDLRVTSKSKESGLSDVSRP
SFCHSSCGHVSEGKRRFTYDHAELERKIACCSPPPDYNSVVLYSTPPI

Protein sequence analysis

PSIPRED

bioinf.cs.ucl.ac.uk/psipred/

Prediction of:

- Secondary structure (also: www.compbio.dundee.ac.uk/www-jpred/)
- Trans-membrane regions (also: www.cbs.dtu.dk/services/TMHMM/)
- Disorder (also: dis.embl.de/)
- Domains
- Function
- 3D structure (homology modelling, fold recognition)

Protein sequence analysis

1) Save Fasta sequence: Ig-like 3, 4, 5
2) Run BLAST
    - Save Sequences of hits in Fasta format
    - Save pair-wise alignments
    - COBALT Multiple Alignment
        - Nb. of sequences
        - Alignment quality

Sequence Comparison Methods (SCM)

2.) Graphic view of matched sequences

Distribution of Blast Hits on the Query Sequence

Sequence Comparison Methods (SCM)

3.) List of matched sequences

Description → pair-wise alignment

Query cover → %age of input sequence matched

E-value → probability that the matched sequence is not homologous

Max ident → % of sequence identity of the longest fragment

Accession → page with protein description

Sequence Comparison Methods (SCM)

4.) Alignments of matched sequences to Query

4.1) ‘Easy’ Results: clear homology

Sequence Comparison Methods (SCM)

4.) Alignments of matched sequences to Query

4.2) ‘Difficult’ Results: homology?

Sequence Comparison Methods (SCM)

Homologous or **Not-homologous?**

1.) % Sequence Identity (%_ID)

2.) Expect value (E-value)

3.) Conservation of key-residues

- e.g., in Ig-like domains: cysteines, tryptophane

Sequence Comparison Methods (SCM)

Homologous or **Not-homologous?**

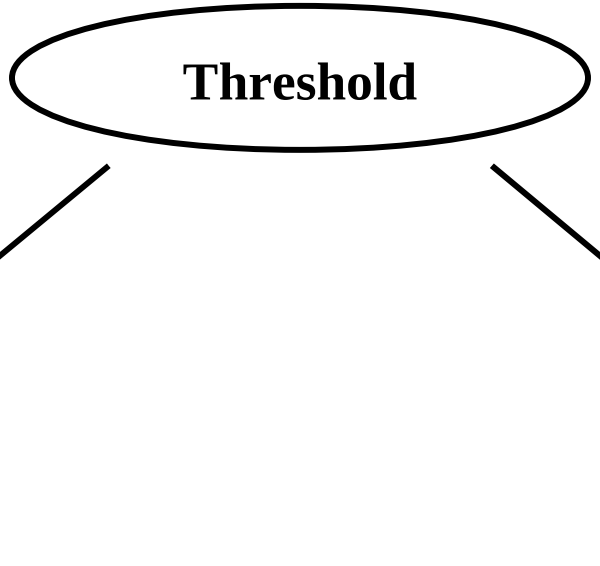
1.) % Sequence Identity (%_ID)

> 30 %

Sequence Comparison Methods (SCM)

Homologous or **Not-homologous?**

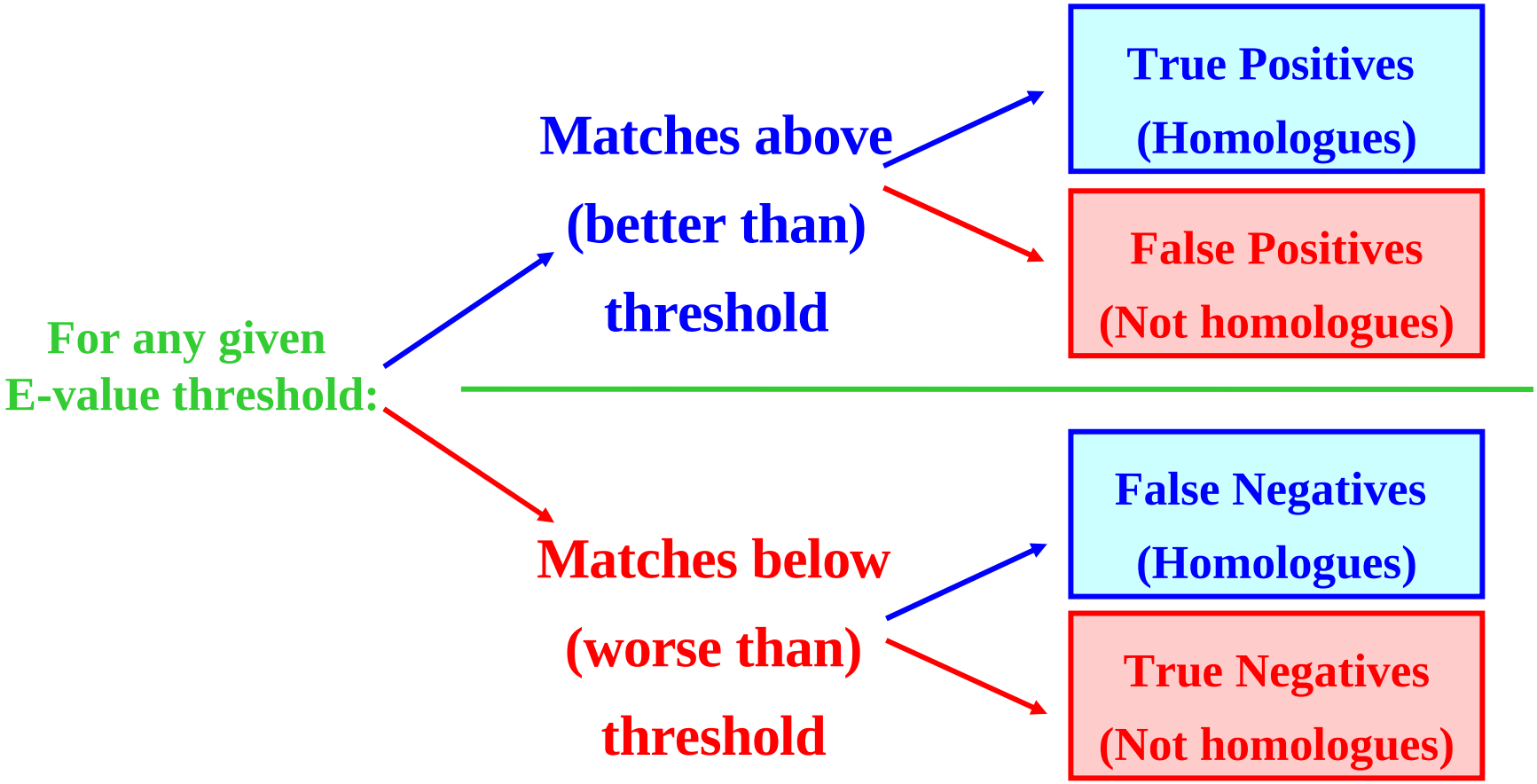
2.) Expect value (E-value): **Number of matches (with a certain score) “expected to be found merely by chance”**



Sequence Comparison Methods (SCM)

Homologous or Not-homologous?

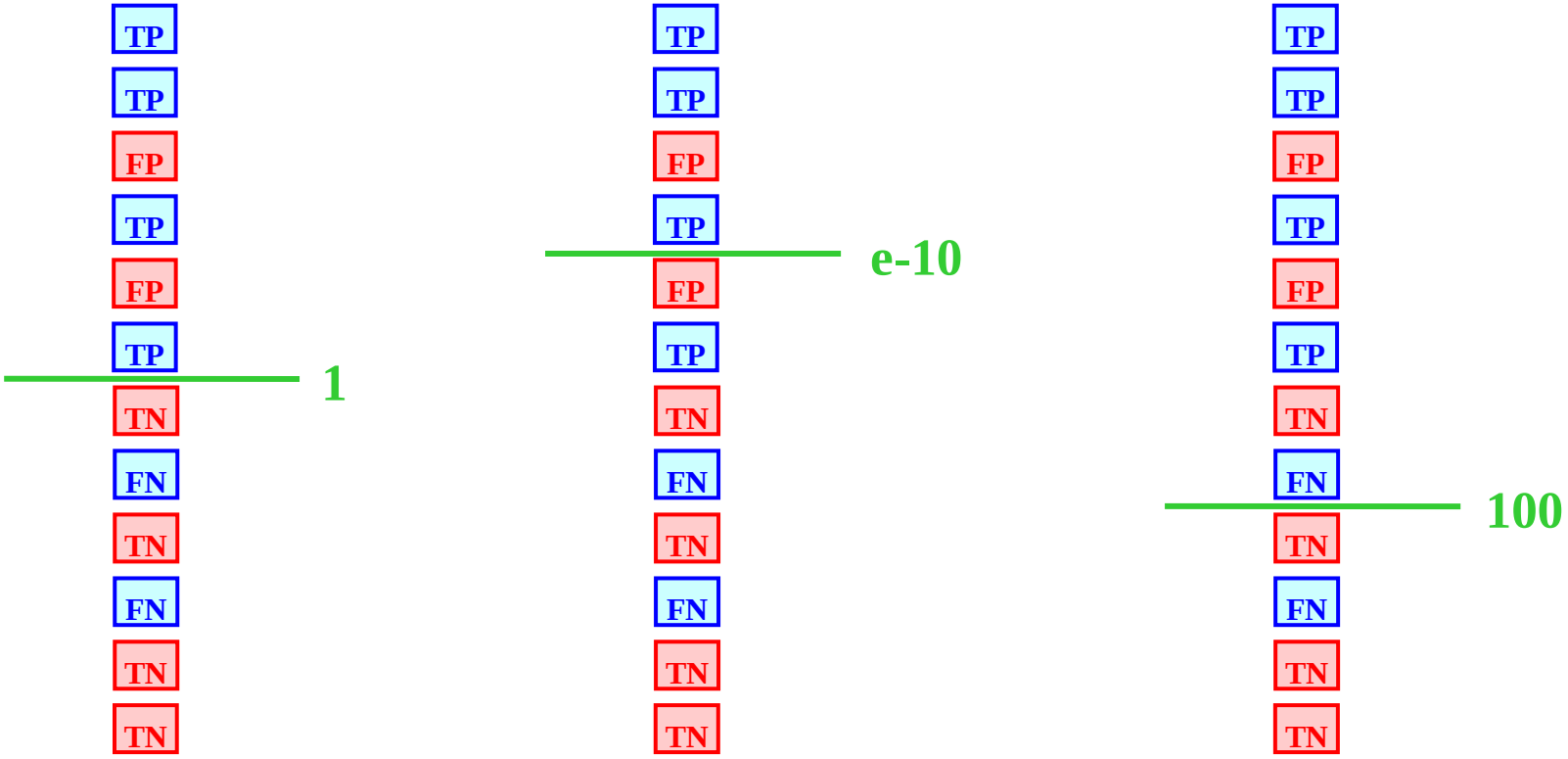
Threshold problem:



Sequence Comparison Methods (SCM)

Homologous or Not-homologous?

2.) Expect value (E-value):



Sequence Comparison Methods (SCM)

Homologous or **Not-homologous?**

1.) % Sequence Identity (%_ID)

2.) Expect value (E-value)

3.) Conservation of key-residues

- MSA
- Literature
- 3D-Structures

Sequence Comparison Methods (SCM)

Pair-wise sequence comparison methods do not recognize “key-residues” for protein structure/function

All positions of the alignment are the same and have the same weight on the computed parameters (i.e., %_ID, E-value, etc.)

Sequence Comparison Methods (SCM)

Multiple sequence alignments (MSA)

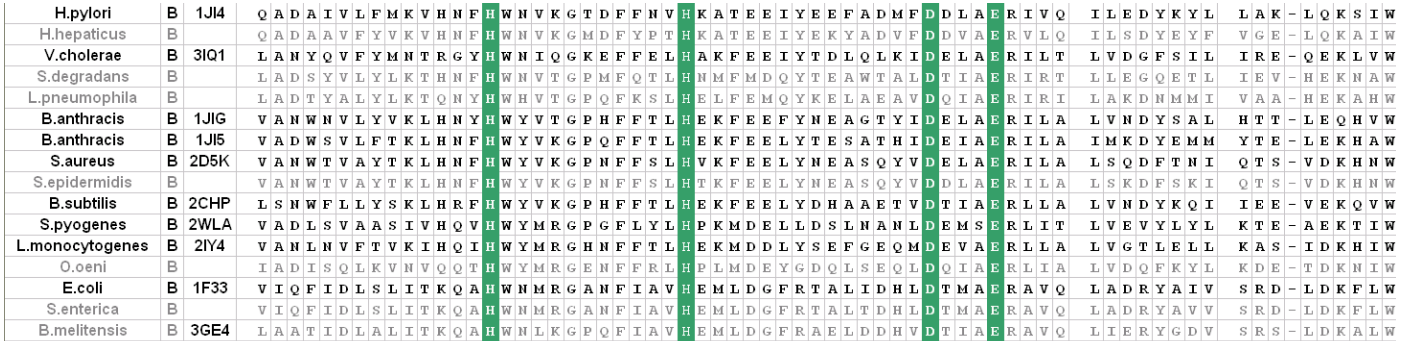
- **More informative than pair-wise alignments**
- **Different positions have different conservation**
- **May allow to recognize “key-residues” for protein structure/function**

- **Input sequences:**
  - **Relatively high number**
  - **Similar enough to produce correct alignments (eliminate ‘outliers’, i.e., < 20 %_ID)**
  - **Different enough to distinguish between conserved and variable positions (make non-redundant, i.e., eliminate > 80 %_ID)**

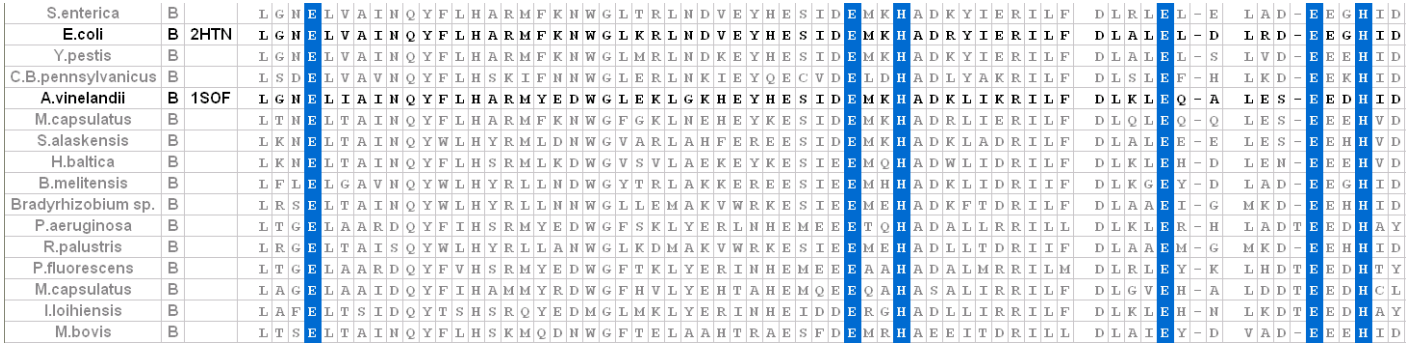
Sequence Comparison Methods (SCM)

‘High-quality’ MSA

Dps proteins



Bacterioferritins



Sequence Comparison Methods (SCM)

Multiple sequence alignment methods

Get sequences to align:

- **putative homologs detected from a Blast search (saved as text)**

Align all sequences in a dataset to:

- **one another**
  - **ClustalW, T-Coffee: www.ebi.ac.uk -> tools -> sequence analysis**

Sequence Comparison Methods (SCM)

Clustal programs

ClustalW2:

- Input sequences
- Multiple Sequence Alignment Options: Aligned vs. Input
- Output (%-age sequence identity)
- Alignment:
    1. Sequences with very different length
    2. Outliers (<20% sequence identity)
    3. Redundant (>80 % sequence identity)

Sequence Comparison Methods (SCM)

Multiple sequence alignment methods

Edit/Visualize MSA:

- **ClustalW/Omega:** www.ebi.org
- **ClustalX:** www.clustal.org; clustalx-2.1-win.msi

- **JalView:** http://www.jalview.org/download.html

- **BioEdit:** http://www.mbio.ncsu.edu/BioEdit/bioedit.html

- **WebLogo:** http://weblogo.berkeley.edu/logo.cgi

Sequence Comparison Methods (SCM)

ClustalW

Input:

- Load sequences
- Step 3: Set your Multiple Sequence Alignment Options

  (Input vs. Aligned)

Output:

- Alignments
- Results Summary: file.output (%ages of sequence identity)

Sequence Comparison Methods (SCM)

ClustalX

Font

File: load sequences

Alignment:

- do complete alignment
- output format options: Clustal vs. Fasta; Input vs. Aligned

Trees: draw tree

Colors

Quality:

- Show low-scoring segments
- Show exceptional residues

Sequence Comparison Methods (SCM)

Bioedit

Graphic view

- Residues per row
- Characters in tiles
- Blocks of ten residues
- Sequences in color
- Outline: similar, identical
- Id/Sim shading
- Id/Sim shading with color table
- Threshold for shading

Sequence Comparison Methods (SCM)

‘High-quality’ MSA

At the basis of a number of structure/function prediction methods:

- domains
- natively unfolded regions
- TM regions
- solvent accessibility
- secondary structures
- 3D-structures

Sequence Comparison Methods (SCM)

Homologous or **Not-homologous?**

1.) % Sequence Identity (%_ID)

2.) Expect value (E-value)

3.) Conservation of key-residues

- **MSA**
- **Literature**
- **3D-Structures**

Sequence Comparison Methods (SCM)

Homologous or **Not-homologous?**

1.) % Sequence Identity (%_ID)

2.) Expect value (E-value)

3.) Conservation of key-residues

- **MSA**
- **Literature**
- **3D-Structures**

Protein structure databases

PDB: Protein Data Bank

www.rcsb.org

PDB Identifier (PDB ID):

- 4 characters: 1st = number; 2nd, 3rd and 4th = letter or number (e.g., 1VFB)

Citation

Molecule description

- Chains, residue numbers

Source

Domain annotation (SCOP)

Protein structure databases

PDB: Protein Data Bank

www.rcsb.org

Method

- X-ray crystallography vs. NMR
- Resolution values

Image - View in 3D

- Mouse options
- Display options: Style (cartoon; backbone; CPK; ball and stick); Color (secondary structure); Surface (solvent accessible); Background; Rotation; S-S bonds; Hydrogen bonds; Export image; etc.

Protein structure databases

PDB: Protein Data Bank

www.rcsb.org

Display files

- Fasta sequence (3 chains)
- PDB file:
  - ATOM: 3rd, atom type; 4th, residue type; 5th, chain name; 6th, residue number; 7th, 8th, 9th: x, y, z co-ordinates; 10th, occupancy; 11th, B-factor
  - TER
  - HETATM

Protein structure databases

PDB: Protein Data Bank

www.rcsb.org

Download files

- Fasta sequence
- PDB file (text)
- Biological Assembly

Sequence

- Secondary structure (DSSP)

Protein structure databases

PDB: Protein Data Bank

www.rcsb.org

SHMT: 1KKJ

- Asymmetric unit vs. Biological assembly (Jmol)
- Ligands and pockets

Protein structure databases

PDBsum

www.ebi.ac.uk/pdbsum/

1VFB

- Protein chains: A, B, C
- Secondary structure, loops, disulfide bonds, catalytic residues, residue conservation

1KKJ

- Protein domains; catalytic residues, PDB sites, contacts to ligands;
- Ligands (ligplot); Clefts (Jmol); Tunnels

Protein annotation databases

Uniprot

www.uniprot.org

Search in

Protein attributes

- Protein existence

General annotation

- Function; Catalytic activity; Subcellular location; …

Sequence annotation

- Amino acid modifications; Variants; …

Protein annotation databases

Uniprot

www.uniprot.org

Cross-references

- 3D structure DBs;
- Protein-protein interactions
  - IntAct: interaction detection method
  - STRING: confidence; evidence; experiments

Sequence Comparison Methods (SCM)

Pairwise methods:

Blast (Fasta; Ssearch)

Profile-based methods:

Psi-Blast (HMMs: SAM-TXX; HMMER)

Profile-profile methods

Sequence Comparison Methods (SCM)

Pairwise methods: Blast (Fasta; Ssearch)

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

(… and mirrors everywhere)

The **Query** sequence is compared

to **each** sequence in a database

Sequence Comparison Methods (SCM)

Pair-wise sequence comparison methods do not recognize "key-residues" for protein structure/function

All positions of the alignment are the same and have the same weight on the computed parameters (i.e., %_ID, E-value, etc.)

How do we overcome this problem?

Structure analysis: best answer / time-consuming, not easy to automatize

Profile-based sequence comparison methods use MSA

Sequence Comparison Methods (SCM)

Profile-based SCM attempt to recognize “key-residues”

Exploit information contained in multiple sequence alignments (i.e., residue conservation in different family members)…

Sequence Comparison Methods (SCM)

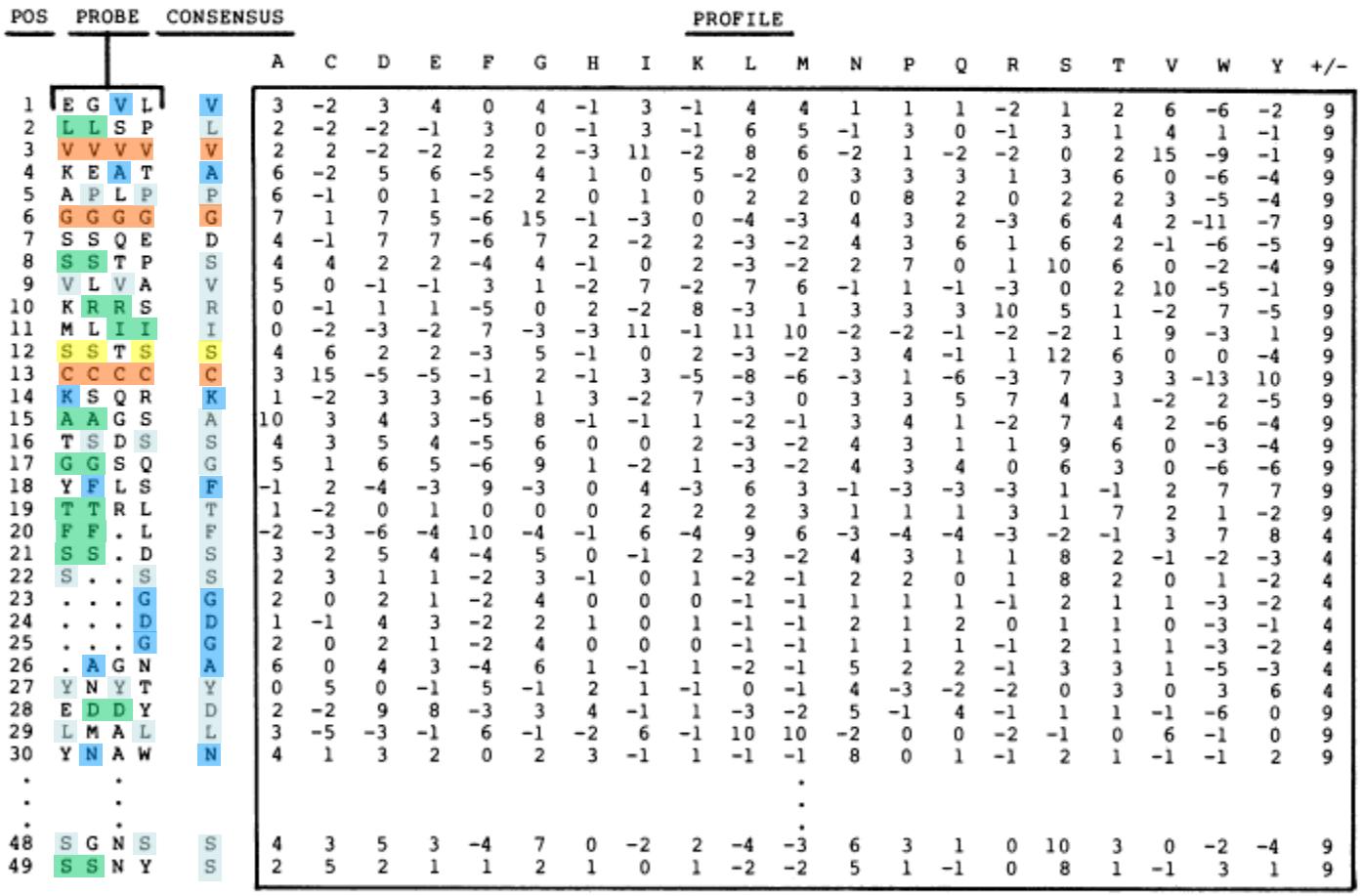
Profile-based SCM attempt to recognize “key-residues”

Exploit information contained in multiple sequence alignments
(i.e., residue conservation in different family members) to build a PROFILE

Query	-	-	-	-	-	M	K	I	H	F	V	G	I	G	G	I	G	M
	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	M
	-	-	M	K	S	L	C	Y	H	F	I	G	I	G	G	I	G	M
	-	-	-	-	-	-	-	I	L	G	L	A	K	S	G	E	A	A
	A	D	Y	Q	G	K	N	V	V	I	I	G	L	G	L	T	G	L
A	0.2																	0.2
C	0.0						0.2											
D	0.0	0.2																
E	0.0																	
F	0.0																	
G	0.0				0.2													
H	0.0																	
I	0.0							0.4										
K	0.0			0.2		0.2	0.2											
L	0.0					0.2												0.2
M	0.0		0.2			0.2												0.6
N	0.0						0.2											
P	0.0																	
Q	0.0			0.2														
R	0.0																	
S	0.0				0.2													
T	0.0																	
V	0.0							0.2										
W	0.0																	
Y	0.0		0.2					0.2										
gaps	0.8	0.8	0.6	0.6	0.6	0.4	0.4	0.2										

Sequence Comparison Methods (SCM)

Consensus vs. PROFILE



POS	PROBE	CONSENSUS	PROFILE																				
			A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	+/-
1	E G V L	V	3	-2	3	4	0	4	-1	3	-1	4	4	1	1	1	-2	1	2	6	-6	-2	9
2	L L S P	L	2	-2	-2	-1	3	0	-1	3	-1	6	5	-1	3	0	-1	3	1	4	1	-1	9
3	V V V V	V	2	2	-2	-2	2	2	-3	11	-2	8	6	-2	1	-2	-2	0	2	15	-9	-1	9
4	K E A T	A	6	-2	5	6	-5	4	1	0	5	-2	0	3	3	3	1	3	6	0	-6	-4	9
5	A P L P	P	6	-1	0	1	-2	2	0	1	0	2	2	0	8	2	0	2	2	3	-5	-4	9
6	G G G G	G	7	1	7	5	-6	15	-1	-3	0	-4	-3	4	3	2	-3	6	4	2	-11	-7	9
7	S S Q E	D	4	-1	7	7	-6	7	2	-2	2	-3	-2	4	3	6	1	6	2	-1	-6	-5	9
8	S S T P	S	4	4	2	2	-4	4	-1	0	2	-3	-2	2	7	0	1	10	6	0	-2	-4	9
9	V L V A	V	5	0	-1	-1	3	1	-2	7	-2	7	6	-1	1	-1	-3	0	2	10	-5	-1	9
10	K R R S	R	0	-1	1	1	-5	0	2	-2	8	-3	1	3	3	3	10	5	1	-2	7	-5	9
11	M L I I	I	0	-2	-3	-2	7	-3	-3	11	-1	11	10	-2	-2	-1	-2	-2	1	9	-3	1	9
12	S S T S	S	4	6	2	2	-3	5	-1	0	2	-3	-2	3	4	-1	1	12	6	0	0	-4	9
13	C C C C	C	3	15	-5	-5	-1	2	-1	3	-5	-8	-6	-3	1	-6	-3	7	3	3	-13	10	9
14	K S Q R	K	1	-2	3	3	-6	1	3	-2	7	-3	0	3	3	5	7	4	1	-2	2	-5	9
15	A A G S	A	10	3	4	3	-5	8	-1	-1	1	-2	-1	3	4	1	-2	7	4	2	-6	-4	9
16	T S D S	S	4	3	5	4	-5	6	0	0	2	-3	-2	4	3	1	1	9	6	0	-3	-4	9
17	G G S Q	G	5	1	6	5	-6	9	1	-2	1	-3	-2	4	3	4	0	6	3	0	-6	-6	9
18	Y F L S	F	-1	2	-4	-3	9	-3	0	4	-3	6	3	-1	-3	-3	-3	1	-1	2	7	7	9
19	T T R L	T	1	-2	0	1	0	0	0	2	2	2	3	1	1	1	3	1	7	2	1	-2	9
20	F F . L	F	-2	-3	-6	-4	10	-4	-1	6	-4	9	6	-3	-4	-4	-3	-2	-1	3	7	8	4
21	S S . D	S	3	2	5	4	-4	5	0	-1	2	-3	-2	4	3	1	1	8	2	-1	-2	-3	4
22	S . . S	S	2	3	1	1	-2	3	-1	0	1	-2	-1	2	2	0	1	8	2	0	1	-2	4
23	. . . G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
24	. . . D	D	1	-1	4	3	-2	2	1	0	1	-1	-1	2	1	2	0	1	1	0	-3	-1	4
25	. . . G	G	2	0	2	1	-2	4	0	0	0	-1	-1	1	1	1	-1	2	1	1	-3	-2	4
26	. A G N	A	6	0	4	3	-4	6	1	-1	1	-2	-1	5	2	2	-1	3	3	1	-5	-3	4
27	Y N Y T	Y	0	5	0	-1	5	-1	2	1	-1	0	-1	4	-3	-2	-2	0	3	0	3	6	4
28	E D D Y	D	2	-2	9	8	-3	3	4	-1	1	-3	-2	5	-1	4	-1	1	1	-1	-6	0	9
29	L M A L	L	3	-5	-3	-1	6	-1	-2	6	-1	10	10	-2	0	0	-2	-1	0	6	-1	0	9
30	Y N A W	N	4	1	3	2	0	2	3	-1	1	-1	-1	8	0	1	-1	2	1	-1	-1	2	9
.	.												.										
.	.												.										
.	.												.										
48	S G N S	S	4	3	5	3	-4	7	0	-2	2	-4	-3	6	3	1	0	10	3	0	-2	-4	9
49	S S N Y	S	2	5	2	1	1	2	1	0	1	-2	-2	5	1	-1	0	8	1	-1	3	1	9

Sequence Comparison Methods (SCM)

Pairwise methods: Blast, Fasta, Ssearch

The **Query** sequence is compared to **each** sequence in a database

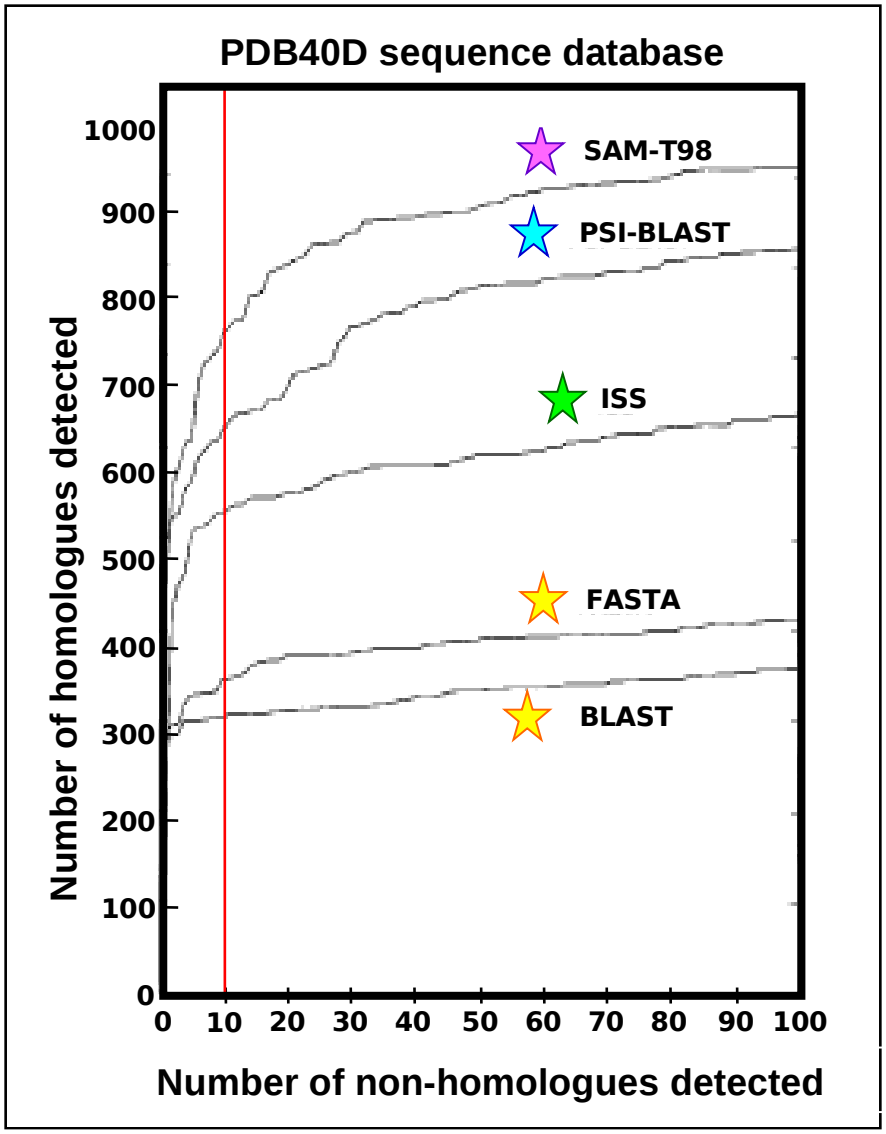
Profile-based methods: Psi-Blast, HMMs

The **Query** sequence is compared to each sequence in a database
The best matches are used to build a **Profile**
The **Profile** is compared to **each** sequence in a database

Profile-profile methods: Psi-Blast, HMMs

The **Query** sequence is compared to each sequence in a database
The best matches are used to build a **Profile**
The **Profile** is compared to the **Profiles** built from each sequence in a database

Sequence Comparison Methods (SCM)



PDB40D sequence database:
sequences of protein domains (D)
of known structure (PDB)
with sequence identity < 40%

Number of homologues detected:
"true positives" (TP)

Number of non-homologues detected:
"false positives" (FP)

Homologues vs. non-homologues:
Structural Classification Of Proteins (SCOP)
database

Pairwise methods

Profile-based methods

Sequence Comparison Methods (SCM)

Profile-based methods: Psi-Blast

Blast, Psi-Blast: http://blast.ncbi.nlm.nih.gov/

(… and mirrors everywhere)

The **Query** sequence is compared to **each** sequence in a database

The best matches are used to build a **Profile**

The **Profile** is compared to **each** sequence in a database

Sequence Comparison Methods (SCM)

Profile-based methods: HMMs

SAM-TXX: http://compbio.soe.ucsc.edu/sam.html

HMMER: http://hmmer.janelia.org/

The **Query** sequence is compared to **each** sequence in a database

The best matches are used to build a **hidden Markov model**

The **hidden Markov model** is compared to **each** sequence in a database

Sequence Comparison Methods (SCM)

Profile-profile methods: HMMs

HHPred: http://toolkit.tuebingen.mpg.de/hhpred

The **Query** sequence is compared to each sequence in a database

The best matches are used to build a **Profile/HMM**

The **Profile/HMM** is compared to the **Profiles/HMM** built from each sequence in a database

Sequence Comparison Methods (SCM)

Pairwise methods (Blast)	→	**Query sequence vs. each DB sequence**
Profile-based methods (Psi-Blast, HMMs)	→	**Profile (built from the query sequence) vs. each DB sequence**
Profile-profile (HMM-HMM) methods	→	**Profile (built from the query sequence) vs. Profile built for each DB sequence**

Sequence Comparison Methods (SCM)

Pairwise methods (Blast)	→	**Query sequence vs. each DB sequence**
Profile-based methods (Psi-Blast, HMMs)	→	**Profile (built from the query sequence) vs. each DB sequence**
Profile-profile (HMM-HMM) methods	→	**Profile (built from the query sequence) vs. Profile built for each DB sequence**

Sequence Comparison Methods (SCM)

Profile-based methods (Psi-Blast, HMMs) → **Profile (built from the query sequence) vs. each DB sequence**

SAM-T08: http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html

Job: http://compbio.soe.ucsc.edu/SAM_T08/results/target08-query-1288251773-1086/summary.html

Metaserver:

- Homology detection
- Secondary Structure Prediction
- Residue-Residue Contact Prediction
- Top-5 models

} "CASP format"

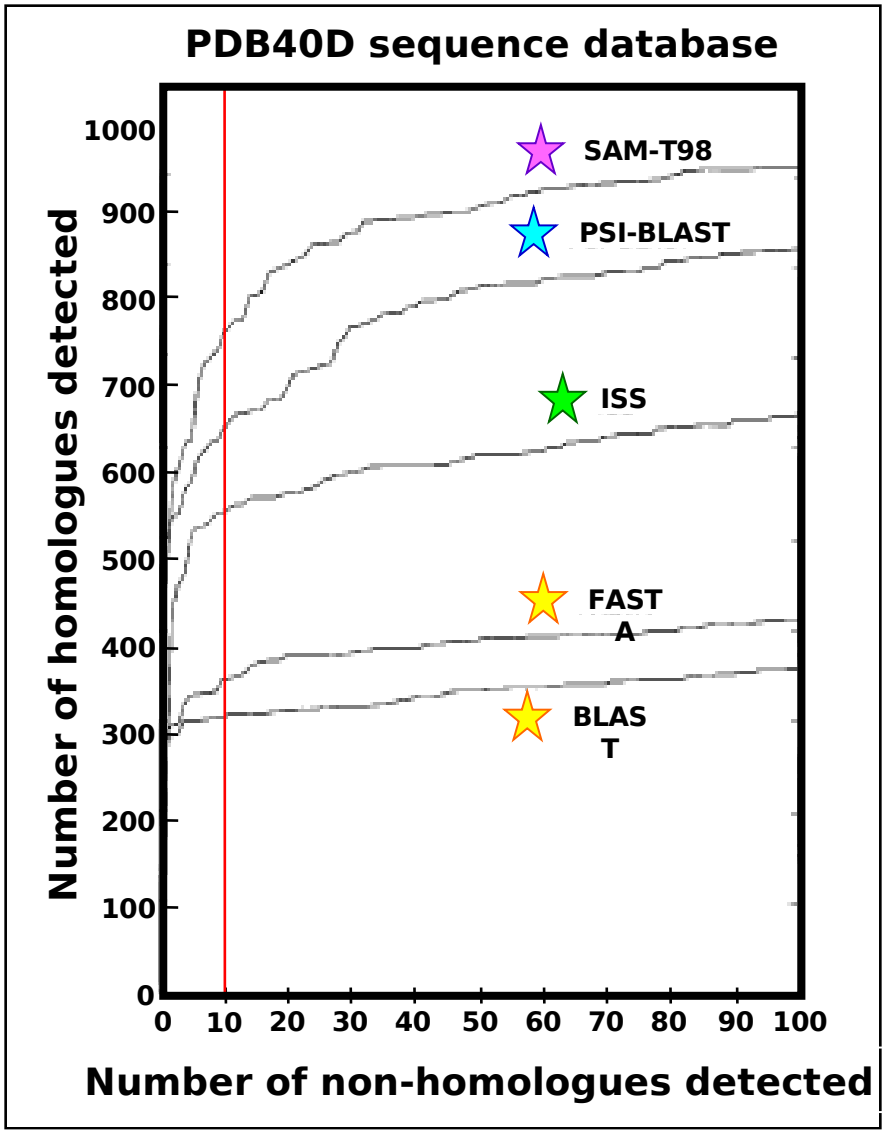
Sequence Comparison Methods (SCM)

Profile-profile (HMM-HMM) methods → **Profile/HMM (built from the query sequence) vs. Profile/HMM built for each DB sequence**

HHpred: http://toolkit.tuebingen.mpg.de/hhpred

Job: http://toolkit.tuebingen.mpg.de/hhpred/results/2645863

Sequence Comparison Methods (SCM)



- Several methods
- Different strategies (sequence-sequence, profile-sequence, profile-profile)
- Similar inputs and outputs
- Different popularity and user-friendliness
- Different ability to recognize distant homologues in performance tests
- BLAST (sequence-sequence)
- PSI-BLAST, SAM-T08 (profile/HMM-sequence)
- HHpred (profile-profile/HMM-HMM)

Sequence Comparison Methods (SCM)

What are they for?

Homology detection

Assignment/Prediction of Structural/Functional properties

- **Detection of a template structure for the whole protein or parts of it**
- **Prediction of protein function and/or functional residues**
- **Prediction of protein architecture (domains, unfolded or transmembrane regions, etc.)**
- **Prediction of promoter regions**

Structure/Function Prediction Methods

How accurate are prediction methods?

For a 3D protein model:

Prediction accuracy:
similarity with the real data

similarity with the real 3D-structure

Accuracy evaluation:
comparison of the prediction with the real data

comparison of the model with the real 3D-structure

If we know the answer (e.g., the real 3D-structure) in advance, can our evaluation be reliable?

We need BLIND TESTS!!!!!!!

Structure/Function Prediction Methods

How accurate are prediction methods?

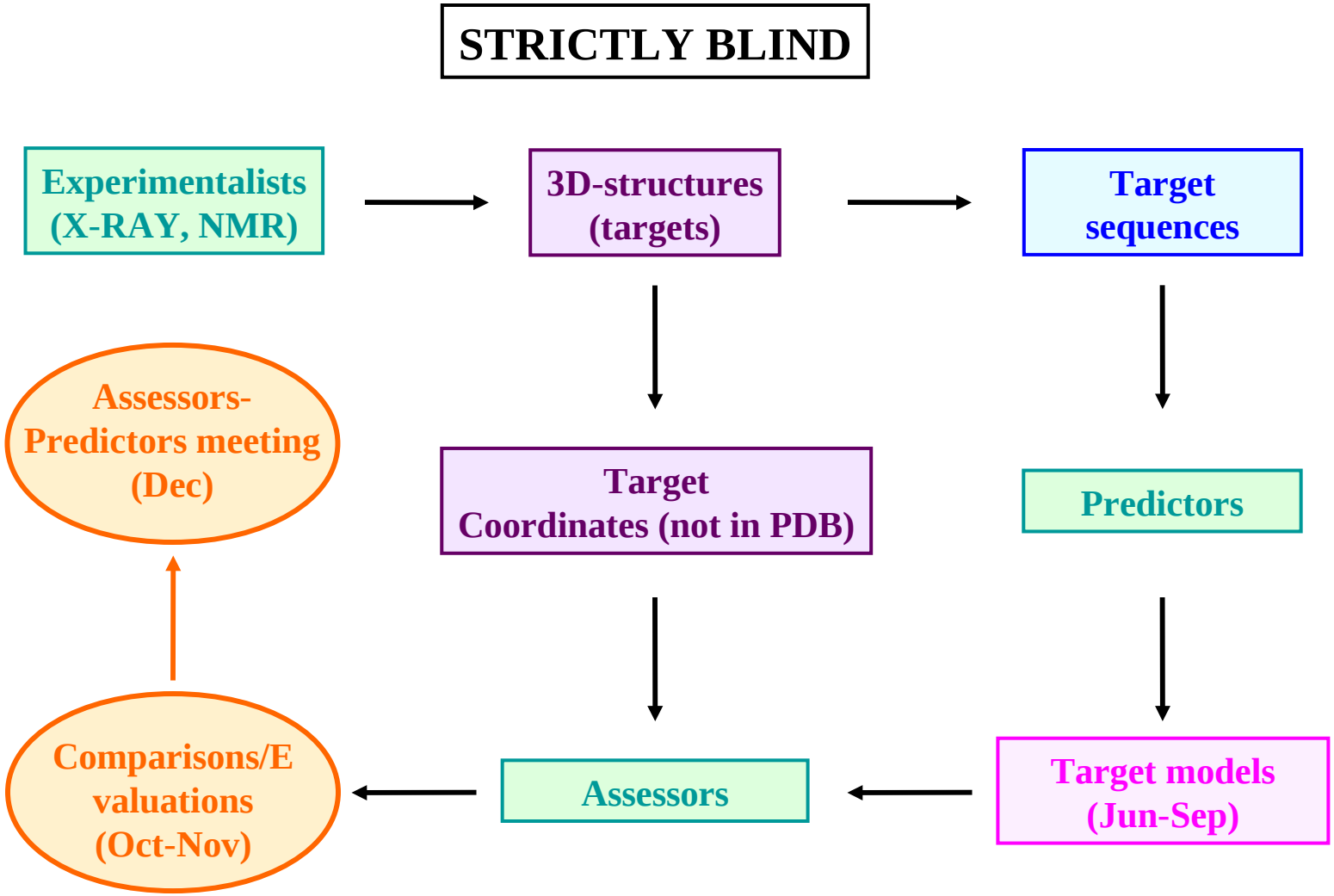
Two types of evaluations

Human-based	Fully automated
Human predictions & Fully automated methods CASP every two years since 1994 CASP9 in 2010	Fully automated methods CAFASP (with CASP) *Livebench, EVA (continuous)*

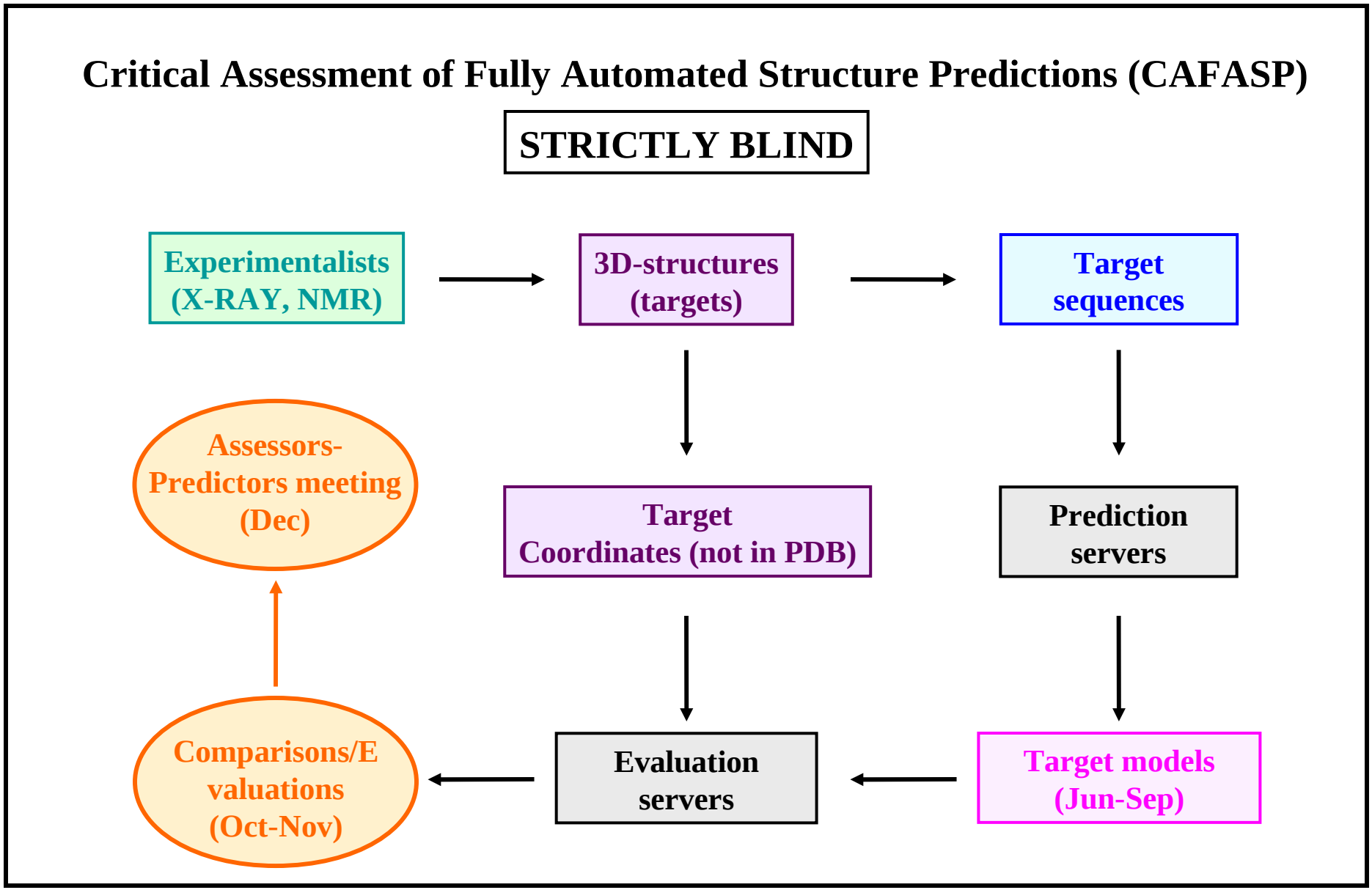
Dramatic performance improvements!!!

Structure/Function Prediction Methods

Critical Assessment of Structure Predictions (CASP)



Structure/Function Prediction Methods



Structure/Function Prediction Methods

CASP, CAFASP RESULTS

- **Reliable picture of the performance of several prediction methods**

Special issues of the Journal:

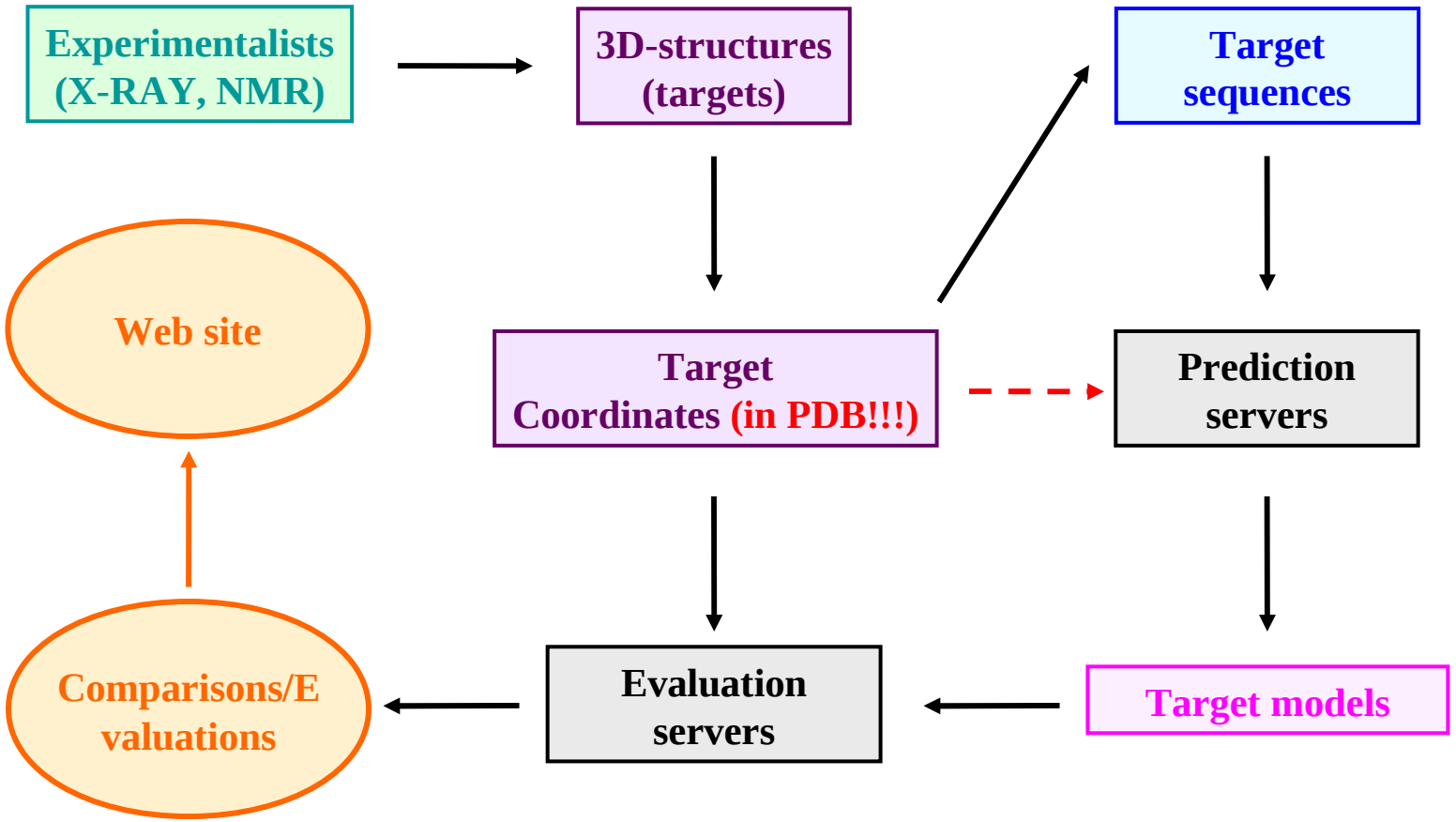
Proteins: Structure, Function and Bioinformatics

- **CASP10 in 2012, special issue in 2013**
- **CASP9 in 2010, special issue in 2011**
- **CASP8 in 2008, special issue in 2009**
- **CASP7 in 2006, special issue in 2007**
- **CASP6 in 2004, special issue in 2005**
- **CASP5 in 2002, special issue in 2003**
- **CASP4 in 2000, special issue in 2001**
- **CASP3 in 1998, special issue in 1999**
- **CASP2 in 1996, special issue in 1997**
- **CASP1 in 1994, special issue in 1995**

Structure/Function Prediction Methods

Livebench, EVA

NOT STRICTLY BLIND: agreed delay of template libraries by 1 week



Structure/Function Prediction Methods

CAPRI: Critical Assessment of Predicted Interactions

- **Performance of protein-protein interaction (docking) methods**

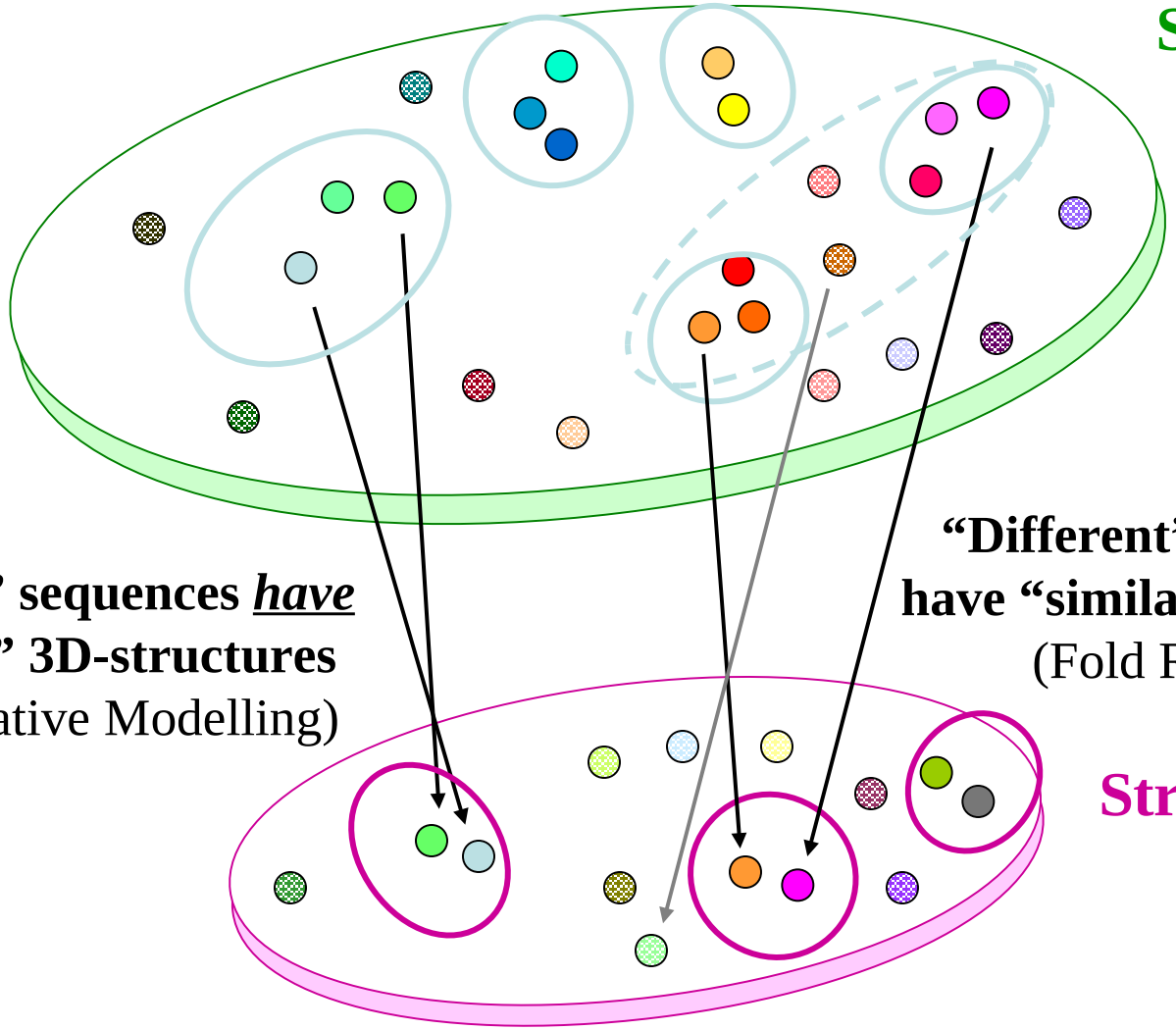
Special issues of the Journal:

Proteins: Structure, Function and Bioinformatics

CAPRI4: current issue

(http://onlinelibrary.wiley.com/doi/10.1002/prot.v78:15/issuetoc)

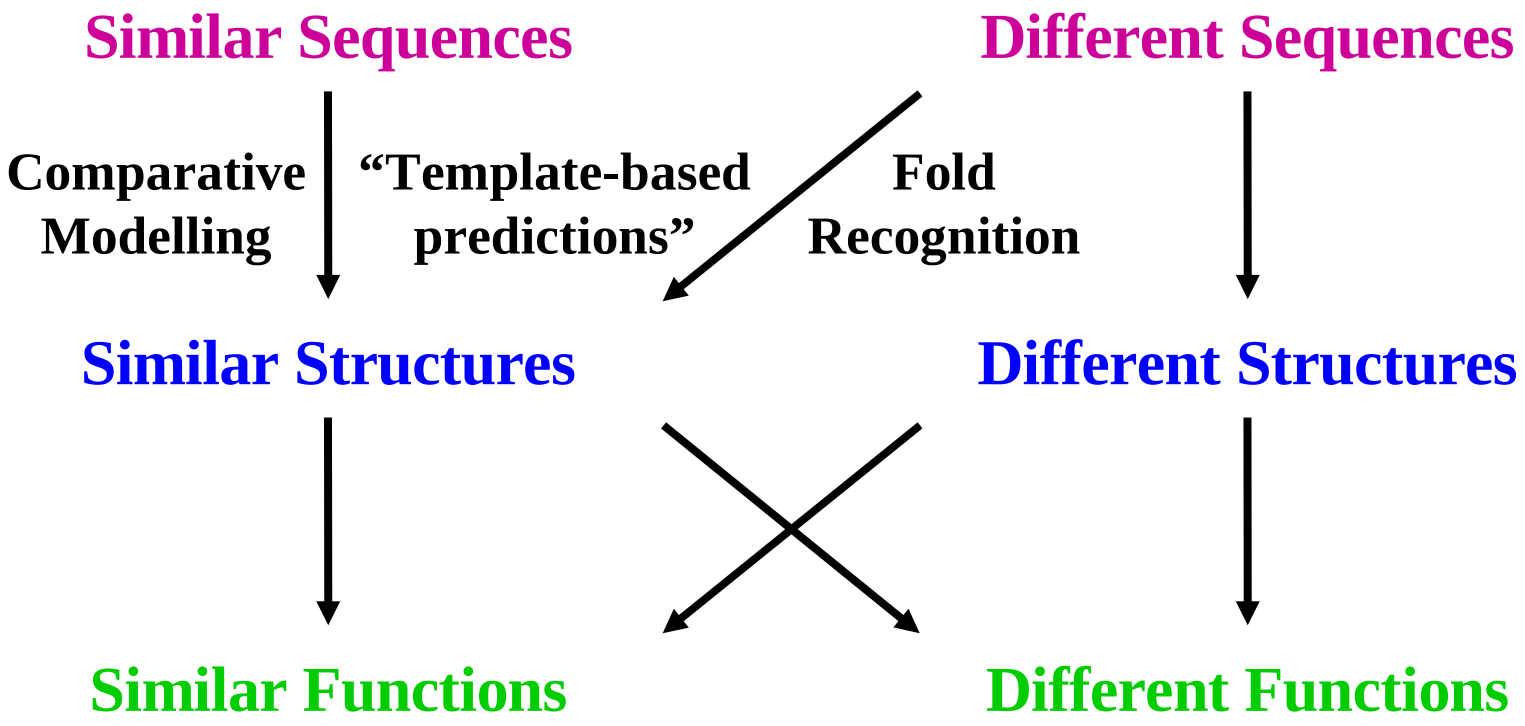
Protein sequence-structure-function relationships



“Similar” sequences *have* “similar” 3D-structures
(Comparative Modelling)

“Different” sequences *can* have “similar” 3D-structures
(Fold Recognition)

Protein sequence-structure-function relationships



Protein sequence-structure-function relationships

a.a. Sequence —?→ 3D-Structure

ab initio
(the "Holy Grail")

Physico-Chemical Principles
(*"nature folds proteins without searching DBs…"*)

evolutionary

Relationships between known sequences and structures (DBs!!!)

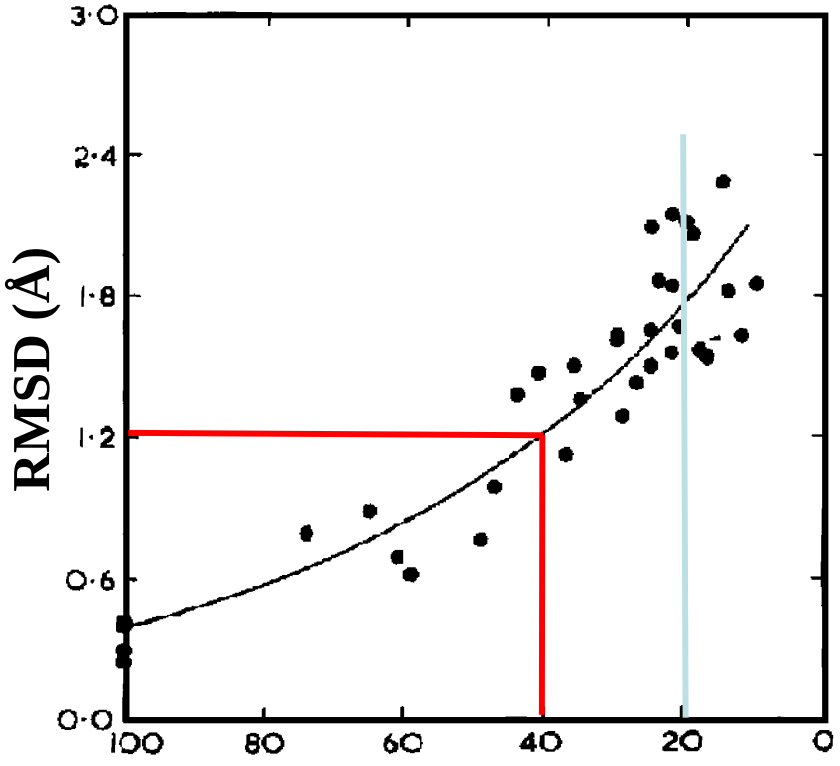
Global level: Template-based predictions

- **High sequence similarity:** **Homology modelling** **++/+++**
- **Low sequence similarity:** **Fold recognition** **+**

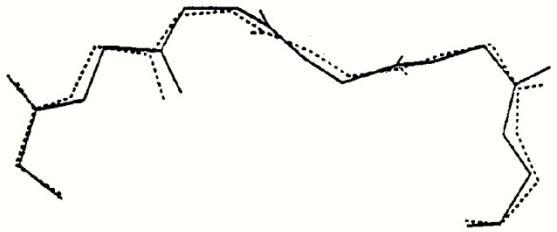
Local level: New-Fold predictions

- **No sequence similarity:** **Fragment-based methods** **-/±**

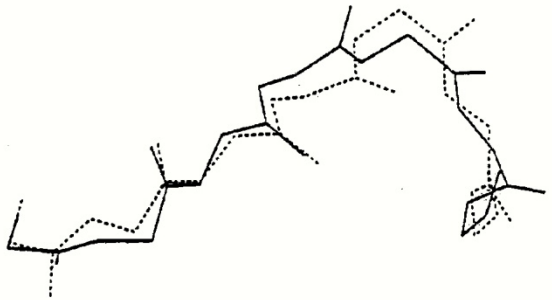
Protein sequence-structure-function relationships



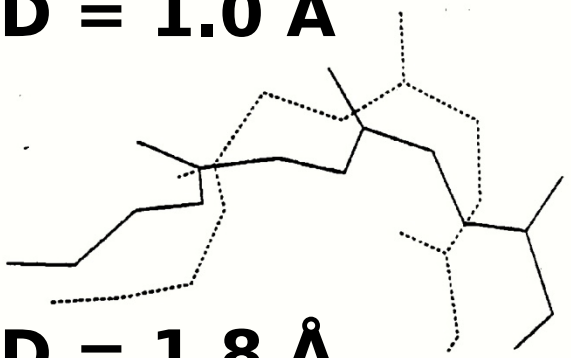
Chothia & Lesk (1986) *EMBO J.*



RMSD = 0.5 Å

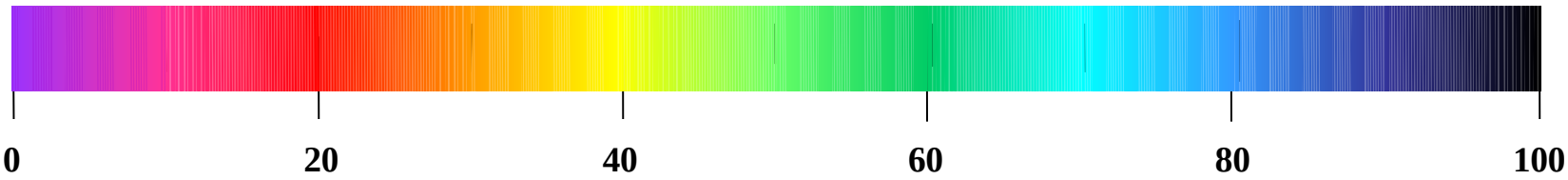


RMSD = 1.0 Å



RMSD = 1.8 Å

Protein structure prediction methods



%age sequence identity with known structures

.....Random.....

Homology modelling

Fold recognition

Template-based: recognition of a global similarity with a protein of known structure

Fragment-based

Global similarity with a protein of known structure not detected: New Fold?

Metaservers + Multiple templates + Model quality evaluation

Protein structure prediction methods

Template-based: Homology modelling

Most accurate => preferred whenever applicable

Procedure:

- **Identify template: protein of known structure homologus to the target (sequence comparison methods).**
- **Produce correct alignment (multiple sequence alignments). Crucial step: errors are inherited in the model.**
- **Identify structurally conserved and variable regions**
- **Replace mutated a.a. in the conserved regions (rotamers)**
- **Model variable regions: 1) alternative templates; 2) loop DBs**
- **Assess reliability: map conserved regions and loops**

Protein structure prediction methods

Template-based: Fold recognition

Less accurate => second choice or as a complement to homology modelling

Procedure like homology modelling, except:

- **Identify template: protein of known structure homologus to the target (fold recognition methods).**

Fold recognition methods: sequence to fold comparisons

- **Target sequence modelled in each structure of a fold representative library (threading): 1D -> 3D**
- **Structures of a fold library described by sequences of structural properties rather than a.a. are compared by SCM to a target sequence described by a sequence of predicted structural properties: 3D -> 1D**

Protein structure prediction methods

New fold: fragment-based

Least accurate => only when all else has failed!

Rational basis:

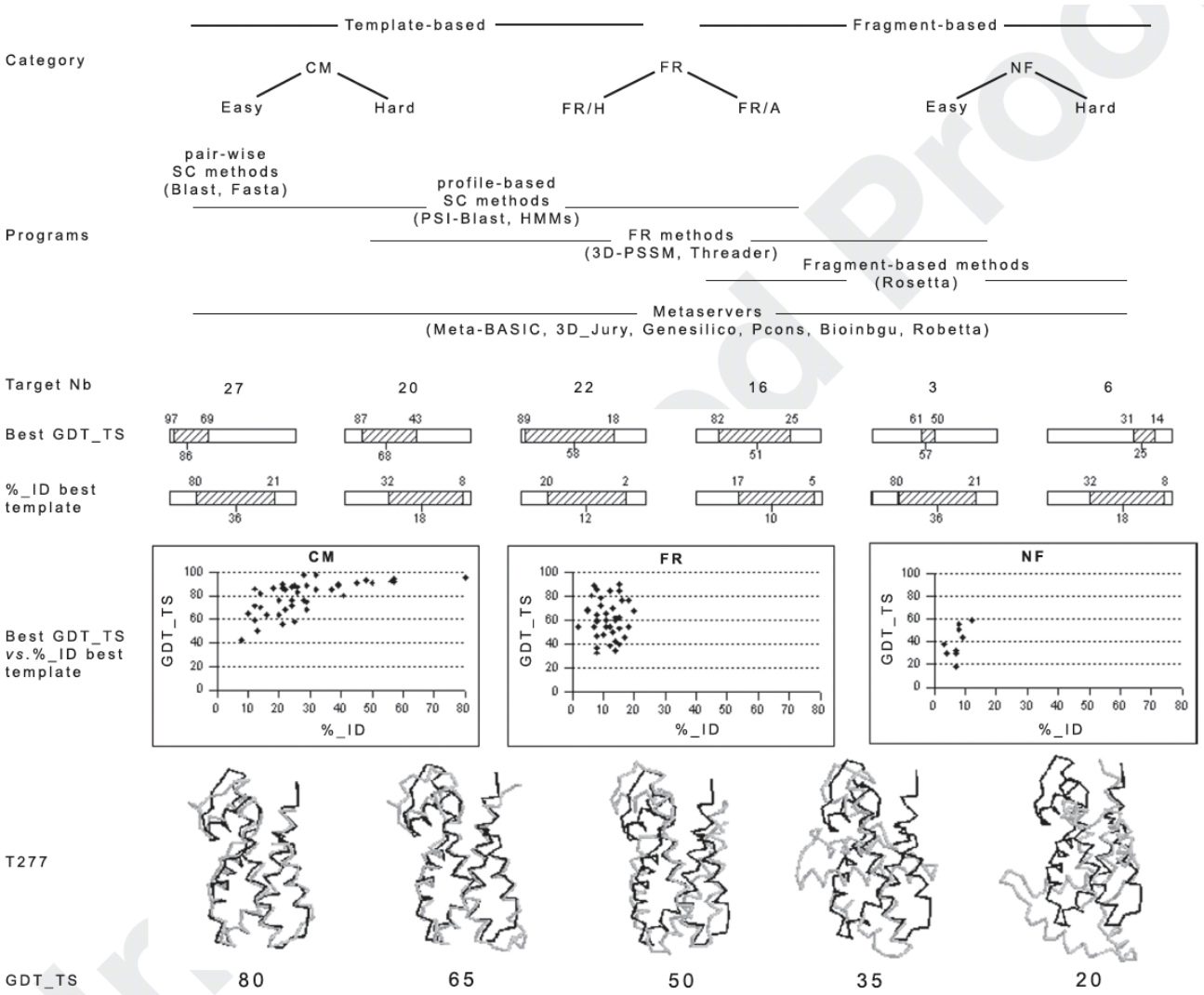
- **Small protein fragments assume a discrete and finite number of conformations**

Procedure:

- **Target sequence is broken into smaller fragments (e.g., 9 and 3 a.a.)**
- **Fragment sequences are used to identify structural fragments with identical sequences => several conformations retrieved for each fragment**
- **Structural fragments (each with many alternative conformations) are combined together to reconstruct the protein fold -> attempt to simulate the folding process**

Protein structure prediction methods

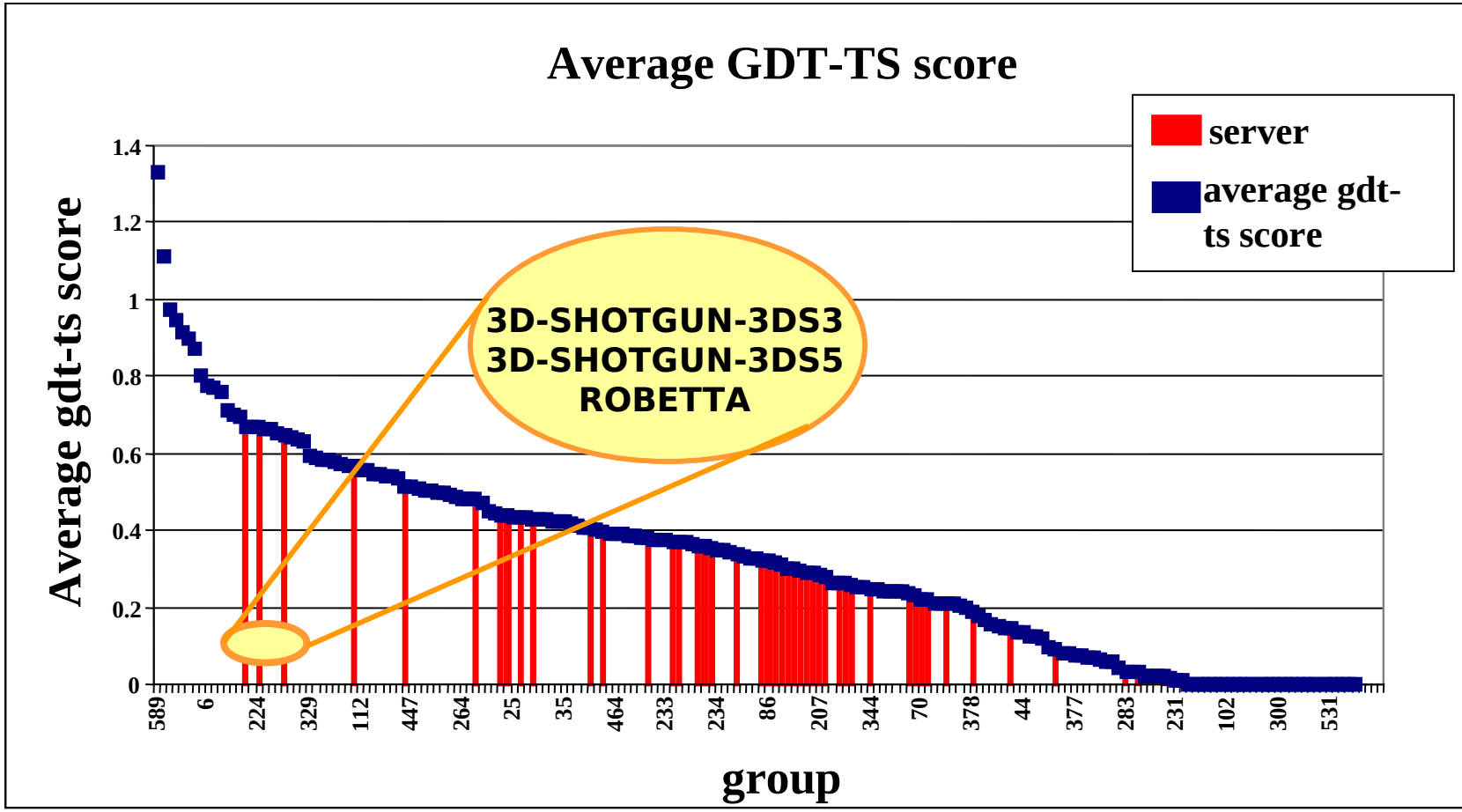
How accurate are prediction methods?



Protein structure prediction methods

How accurate are prediction methods?

Human experts perform better than automated methods



Protein structure prediction methods

What is the purpose of the model?

Procedure to choose

Required Accuracy
Vs.
Time available

Consider experimental (X-ray, NMR, EM) structure determination!

Protein structure prediction methods

What is the purpose of the model?

Biological applications of protein structure prediction methods

High accuracy 3D model:

- **drug design; docking**

General model at the fold level:

- **function prediction**

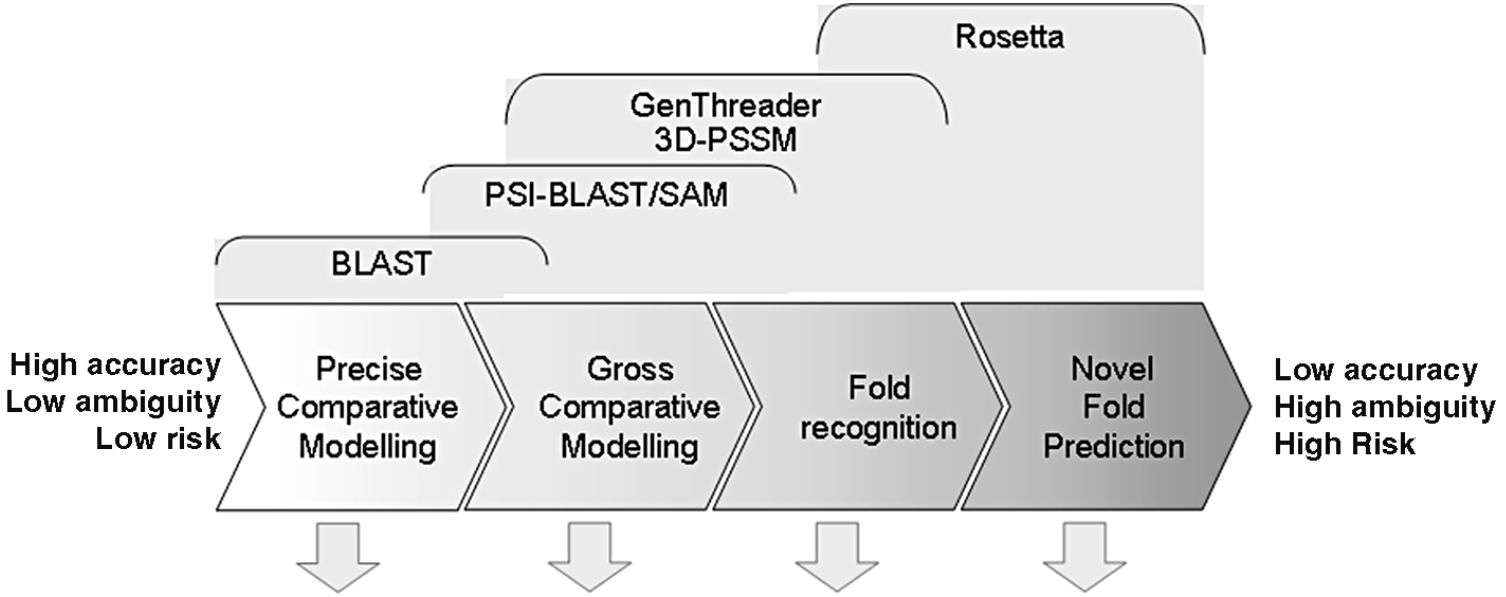
Topology / Globular vs. Natively unfolded:

- **engineer insoluble proteins into smaller and soluble portions**

Protein structure prediction methods

What is the purpose of the model?

Biological applications of protein structure prediction methods



Drug Design, Docking	+++			
SAR rationalization	+++	++		
Construct design	+++	+++		
Function prediction	+++	+++	++	+
Assay selection	+++	++	+	

Protein structure prediction methods

How do we proceed?

- **Fully automated methods and Model DBs**
  - **Modeller/ModBase**
  - **Swiss-Model/Swiss-Model Repository**
  - **3D-Jury, 3D-shotgun, Pcons, Pmodeller *(Genesilico, Meta-BASIC)* (CM, FR)**
  - **Robetta (FR, NF)**
- **Semi-automated: produce the alignment, use program to build the model (transform the alignment in 3D coordinates)**
  - **Modeller**
- **Manual**

Protein structure prediction methods

Secondary structure prediction methods

Tools for prediction of:

- **Domains**
- **Disordered regions**
- **Trans-membrane regions**
- **Secondary structure elements**