新時代の計算生物学

産業技術総合研究所

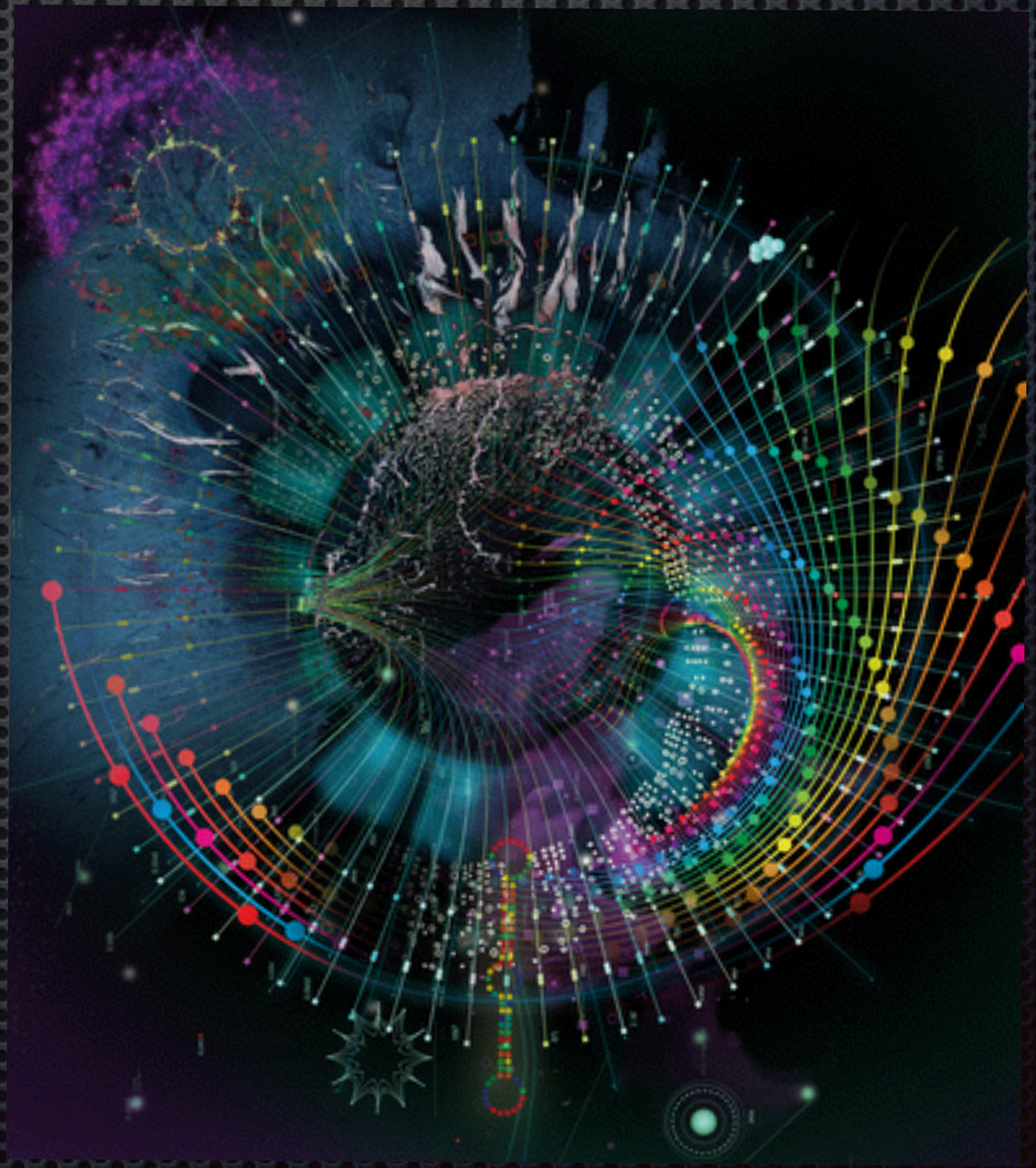
富井 健太郎

第8回「学際計算科学による新たな知の発見・統合・創出」シンポジウム

平成28年10月17日

Outline

- 計算生物学とは?
- アラインメント
  - アミノ酸置換行列の改良
  - マルチプルアラインメント
- Deep Learning
  - 二次構造予測



Tatiana Plakhova

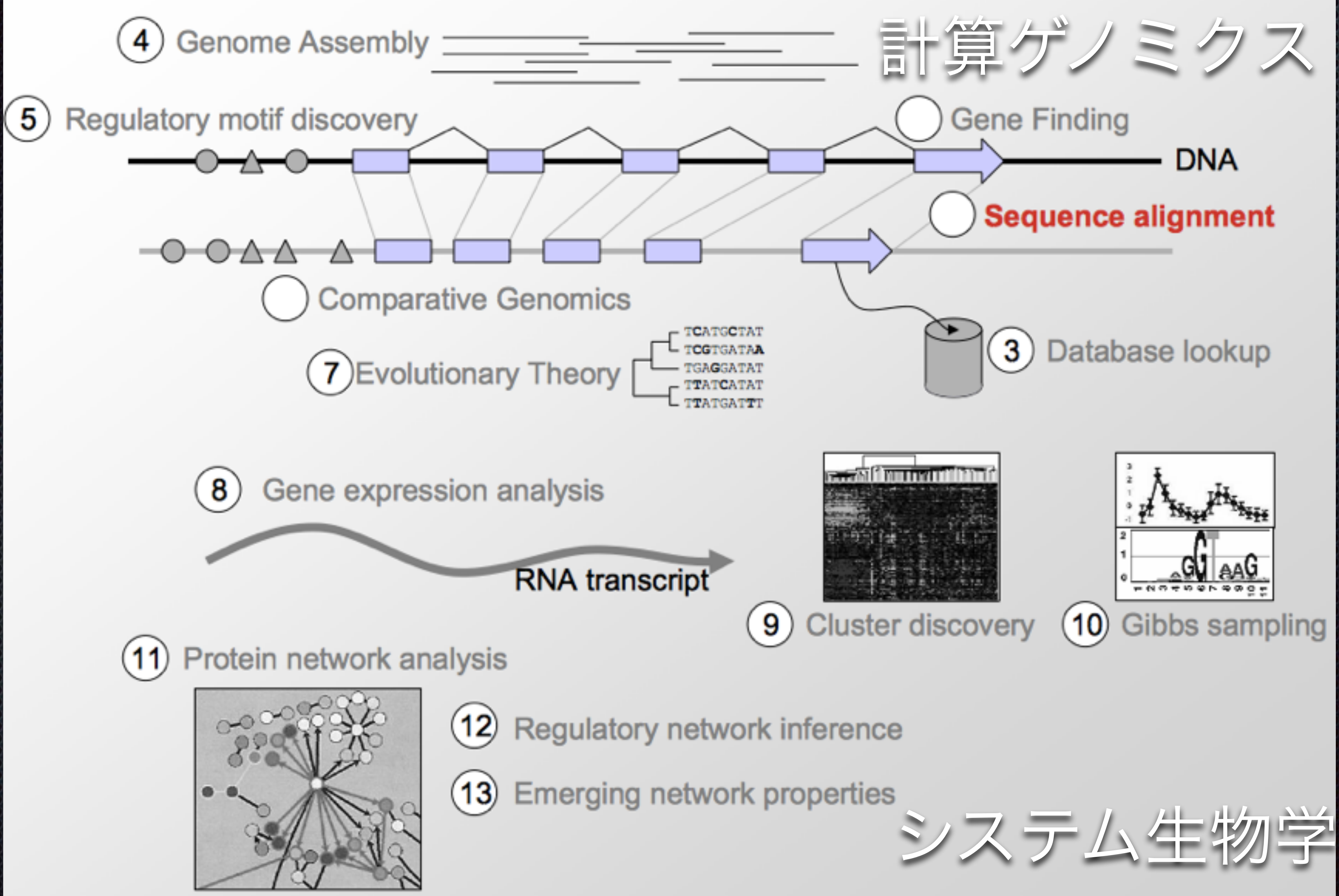
Nature **527**, S2–S4 (05 November 2015)

計算生物学(Computational Biology)

- 生物学の問題の解決に計算機科学、応用数学、統計学の手法を応用する学際研究分野。
  - バイオインフォマティクス (Bioinformatics)
  - 計算生物モデリング
  - 計算ゲノミクス
  - 分子モデリング
  - システム生物学
  - タンパク質構造予測と構造ゲノミクス
  - 計算生化学と計算生物物理学

ja.wikipedia.org

Challenges in Computational Biology

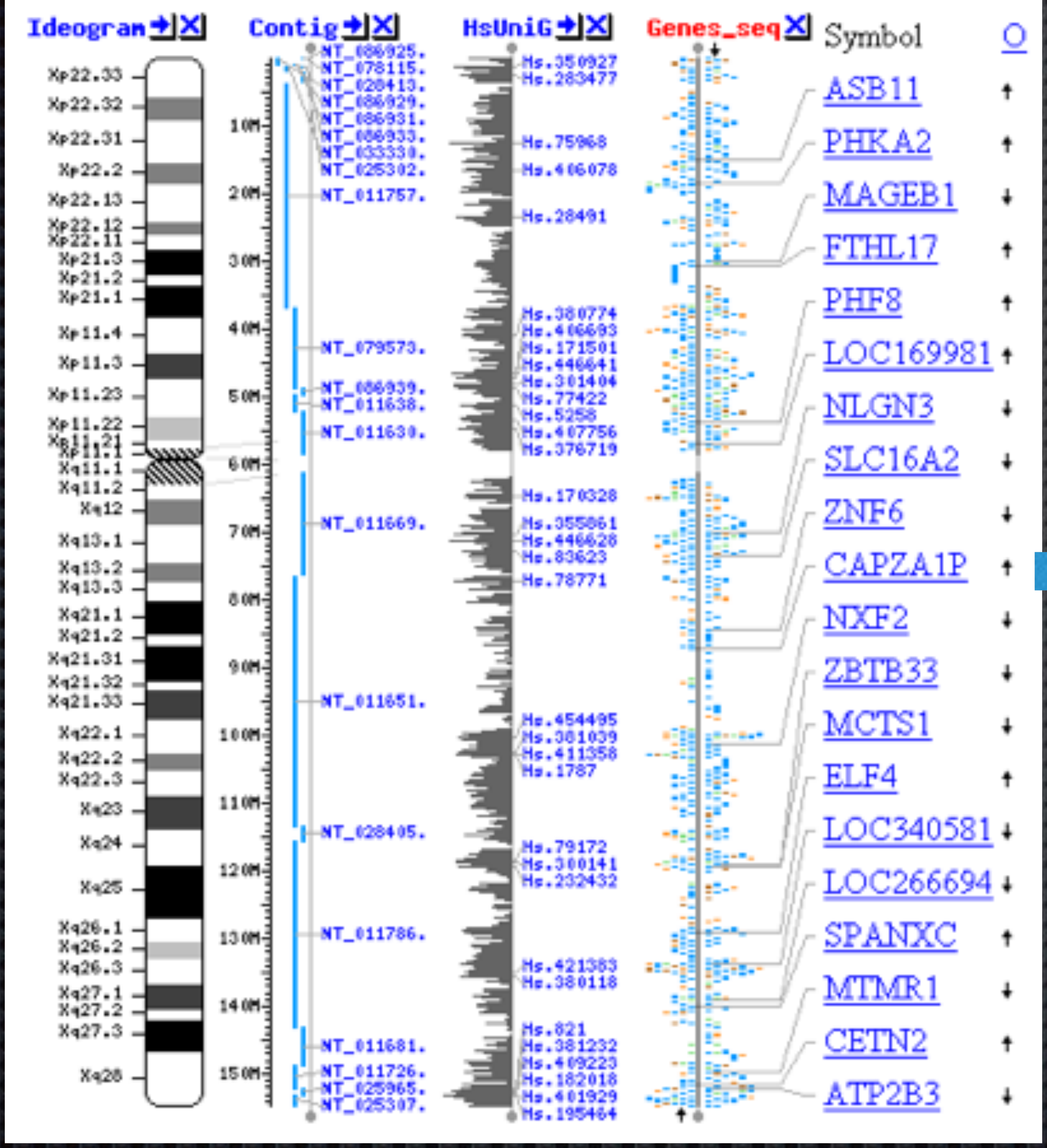


計算ゲノミクス

システム生物学

Algorithms for Computational Biology ocw.mit.edu/courses/

バイオインフォマティクス/
生命情報学



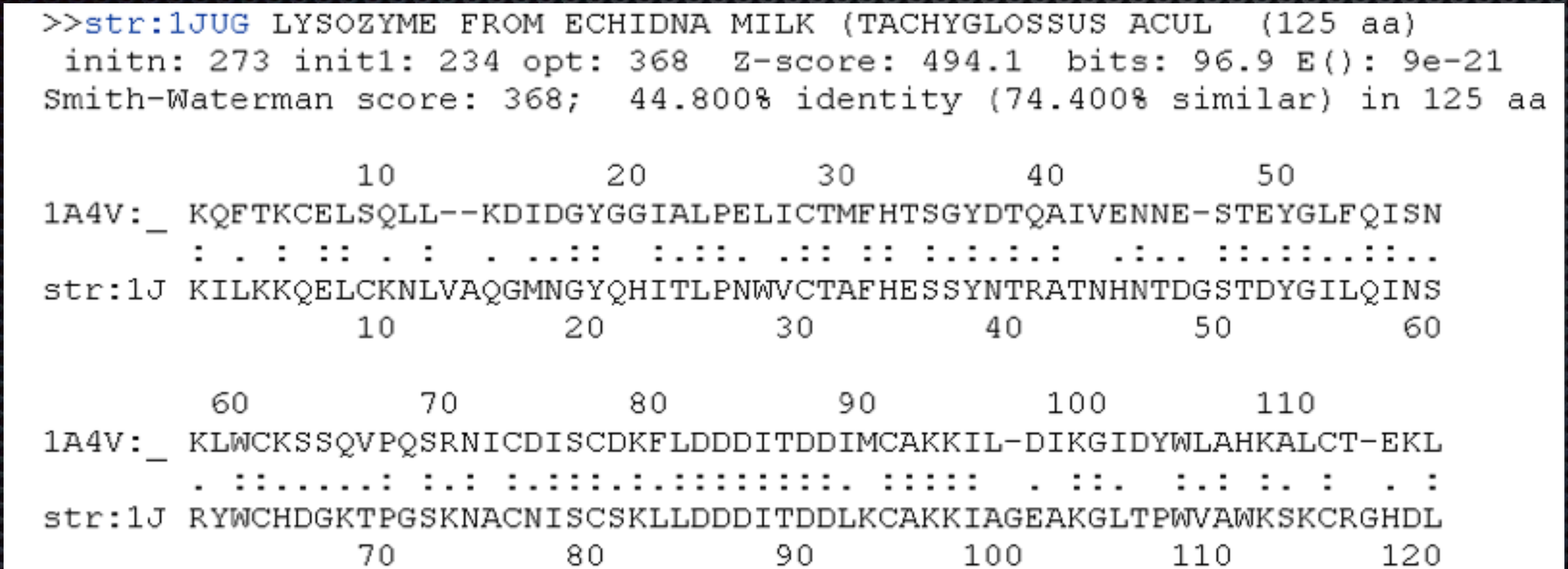
ヒトX染色体 Genome viewer screenshot

• • • • • • • • • • • • • • •

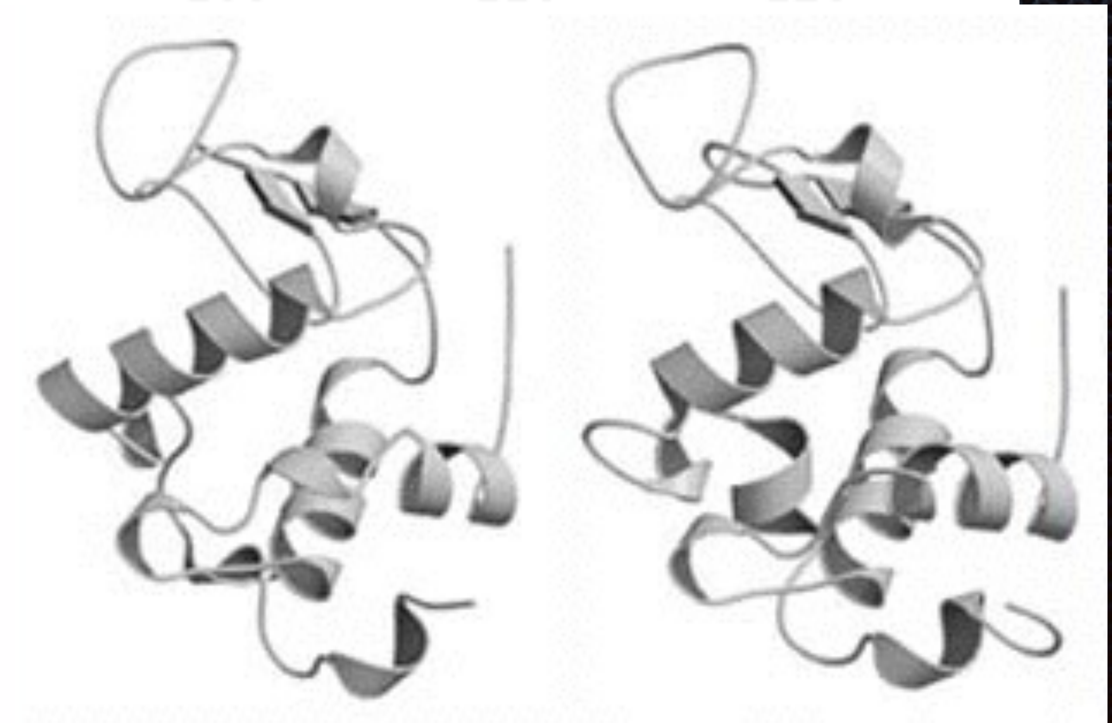
TATTTTGTTTTTGAAAGCCAGTAAA
TTTGTATTAATATCTCATGGCTAGA
GTTCTGAAGTAAAAGTTACAGAATT
TGTGTGTGTGAGTGTGTGTGTGTTT
GTGTGTGTATATATTTAAAAGGCCT
TTATGATAGATTTCTATTTTATGTT
TAAATGGCAATTAAGCTGGTTTTGA
TTTCCCTCTAGCACACCAGACTTTT
TCTCTCTTTACTTTGAGATGTACGT
TTTTGTTATCTAATTTTTCACCTAA
GGGTTATTTTCTTCAATATGAAAAT
TTGTGGTTATTTAGCTGACAATTAC
CTAGGGTAATAAAATAGGTTATCAT
TTTGAAAGTGTGAAAAAAAGGTCTT

• • • • • • • • • • • • • • •

similar proteins

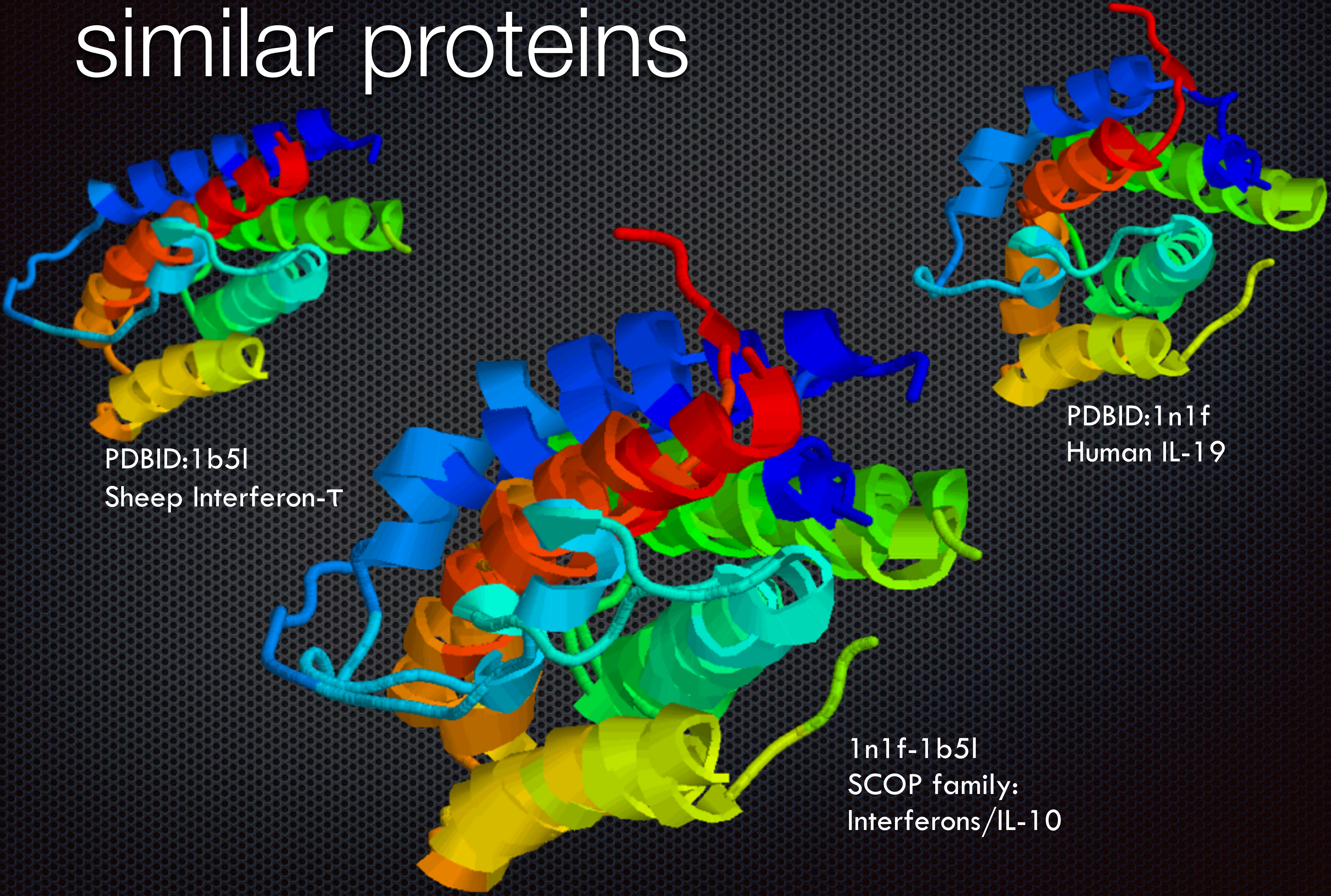


```
>>str:1JUG LYSOZYME FROM ECHIDNA MILK (TACHYGLOSSUS ACUL  (125 aa)
 initn: 273 init1: 234 opt: 368  Z-score: 494.1  bits: 96.9 E(): 9e-21
Smith-Waterman score: 368;  44.800% identity (74.400% similar) in 125 aa

               10          20        30        40         50
1A4V:_ KQFTKCELSQLL--KDIDGYGGIALPELICTMFHTSGYDTQAIVENNE-STEYGLFQISN
       : . : :: . :  . ..::  :.::. .:: :: :.:.:.:  .:.. ::.::..::..
str:1J KILKKQELCKNLVAQGMNGYQHITLPNWVCTAFHESSYNTRATNHNTDGSTDYGILQINS
               10        20        30        40        50        60

        60        70        80        90        100       110
1A4V:_ KLWCKSSQVPQSRNICDISCDKFLDDDITDDIMCAKKIL-DIKGIDYWLAHKALCT-EKL
       . ::.....: :.: :.:::.:.::::::::. :::::  . ::.  :.: :. :  . :
str:1J RYWCHDGKTPGSKNACNISCSKLLDDDITDDLKCAKKIAGEAKGLTPWVAWKSKCRGHDL
               70        80        90       100       110       120

         120
1A4V:_ EQWLCEKL
        .. :
str:1J SKFKC
```



ラクトアルブミン　　　リゾチーム

similar proteins



PDBID:1b5l
Sheep Interferon-τ

PDBID:1n1f
Human IL-19

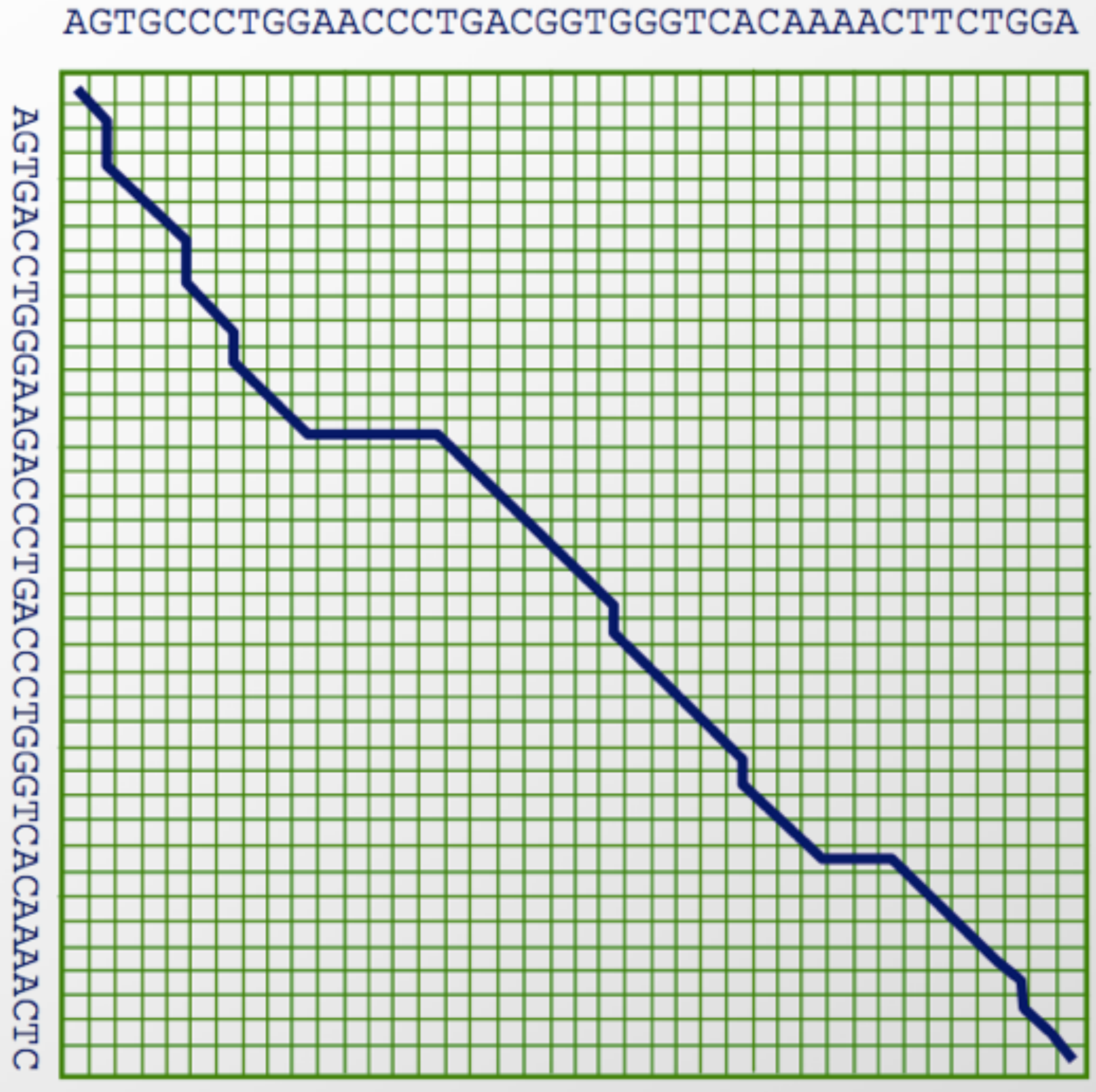
1n1f-1b5l
SCOP family:
Interferons/IL-10

pairwise alignment

- 1N1F:A(size=159) vs 1B5L:_(size=172)
- Structure Alignment Rmsd = 3.0Å, Z-Score = 5.3
- Sequence identity = 8.1% (11/136)
- Aligned/gap positions = 136/25
- Sequence alignment based on structure alignment by CE (cl.sdsc.edu).

```
1N1F:A   ISTDMHHIEESFQEIKRAIQAKDTFPNVTILSTLETLQII--------KPLDVCCVTKNL
1B5L:_   LMLDARENLKLLDRMNRLSPHSCLQDRKDF-GL--PQEMVEGDQLQKDQAFPVLYEMLQQ


1N1F:A   LAFYVDRVFKDHQEPNPKILRKISSIANSFLYMQKTLRQCQEQRQCHC------RQEATN
1B5L:_   SFNLFYTEHSSAAWD----TTLLEQLCTGLQQQLDHLDTCRGQVMGEEDSELGNMDPIVT


1N1F:A   ATRVIHDNYDQ---LEVHA-AAIKSLGELDVFLAWINKNHE
1B5L:_   VKKYFQGIYDYLQEKGYSDCAWEIVRVEMMRALTVSTTLQK
```

How do we compute the best alignment?


AGTGCCCTGGAACCCTGACGGTGGGTCACAAAACTTCTGGA
AGTGACCTGGGAAGACCCTGACCCTGGGTCACAAAACTC

Too many possible alignments:

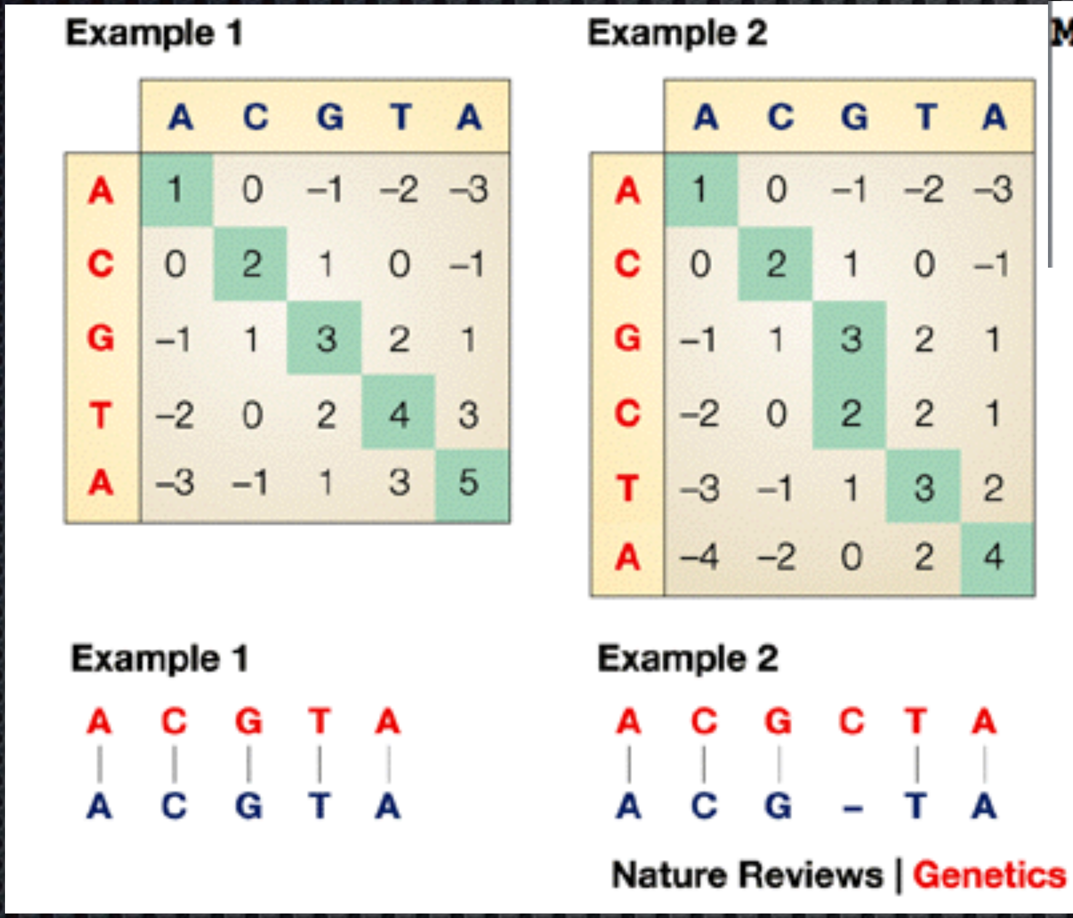
$$O(\ 2^{M+N})$$

Ways to align two sequences of length m, n

$$\binom{n+m}{m} \quad \frac{(m+n)!}{(m!)^2} \approx \frac{2^{m+n}}{\sqrt{\pi m}}$$

Algorithms for Computational Biology ocw.mit.edu/courses/

How do we compute the best alignment?



Example 1

	A	C	G	T	A
A	1	0	−1	−2	−3
C	0	2	1	0	−1
G	−1	1	3	2	1
T	−2	0	2	4	3
A	−3	−1	1	3	5

Example 2

	A	C	G	T	A
A	1	0	−1	−2	−3
C	0	2	1	0	−1
G	−1	1	3	2	1
C	−2	0	2	2	1
T	−3	−1	1	3	2
A	−4	−2	0	2	4

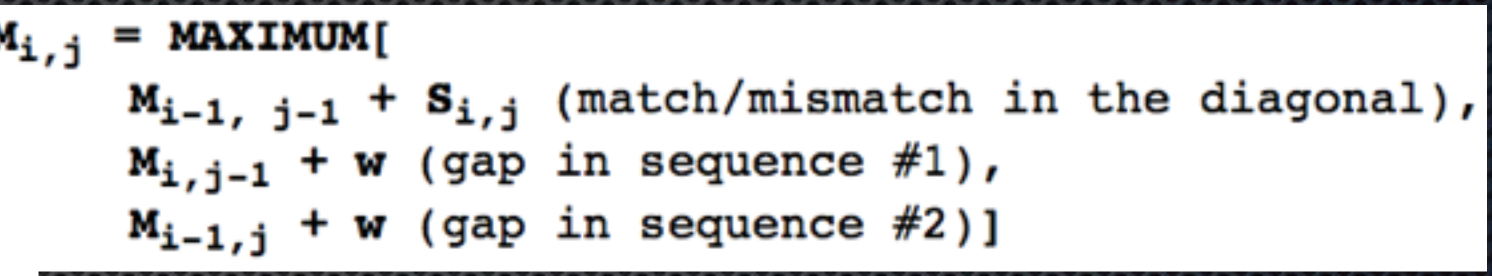
Example 1

```
A C G T A
| | | | |
A C G T A
```

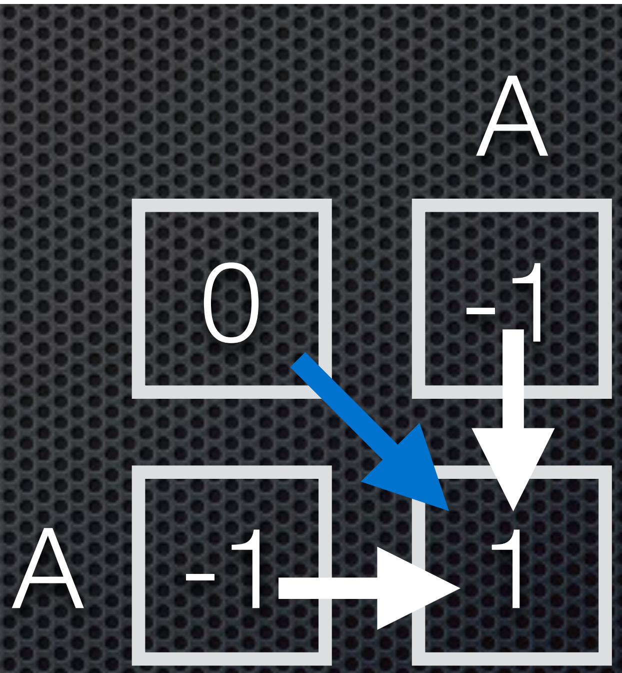
Example 2

```
A C G C T A
| | |   | |
A C G – T A
```

Nature Reviews | Genetics



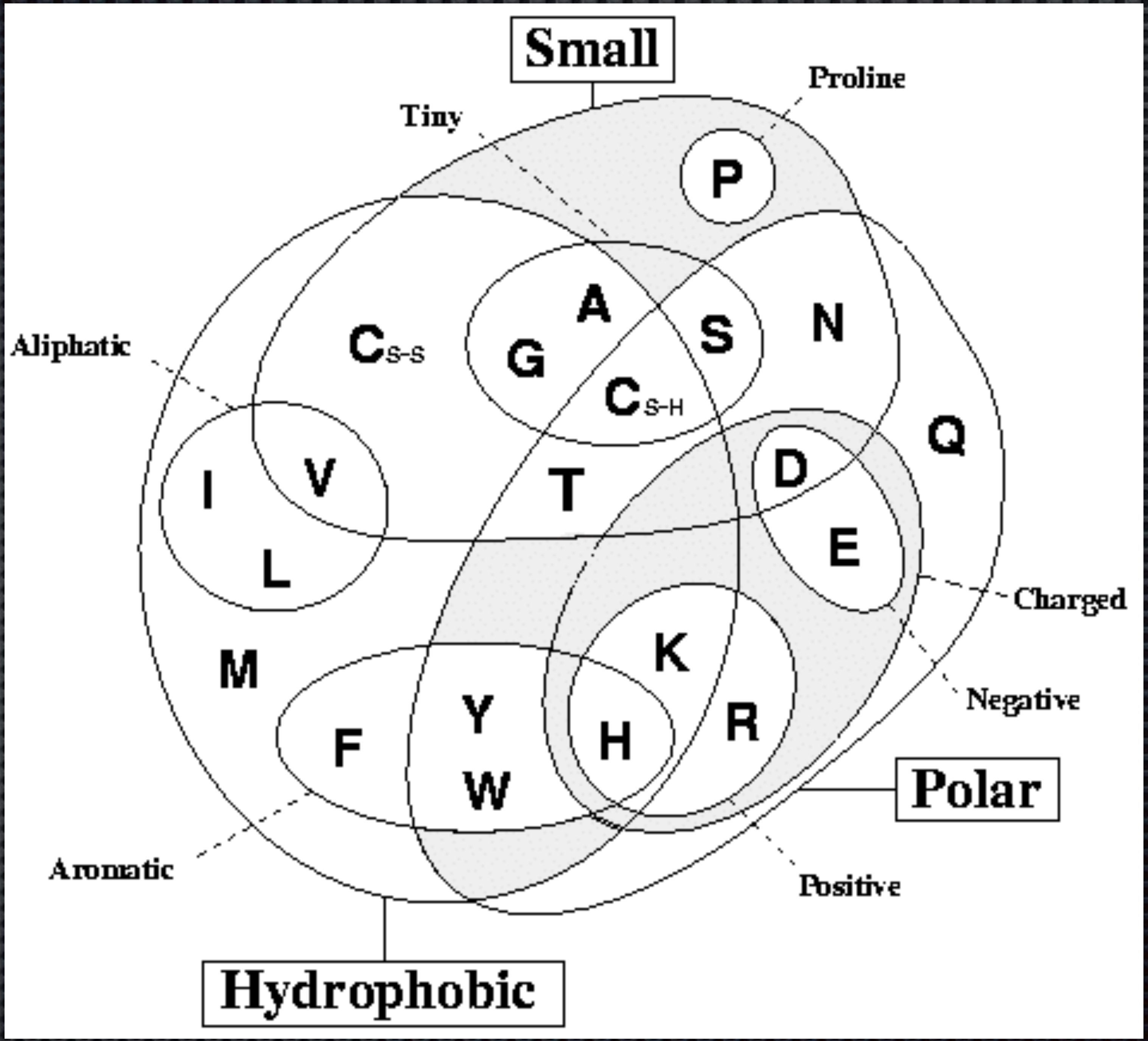
```
Mi,j = MAXIMUM[
     Mi-1, j-1 + Si,j (match/mismatch in the diagonal),
     Mi,j-1 + w (gap in sequence #1),
     Mi-1,j + w (gap in sequence #2)]
```



a match is scored as 1
a mismatch is scored as -1
an insertion/deletion gap penalty is scored as -1

Amino acid properties

脂肪族

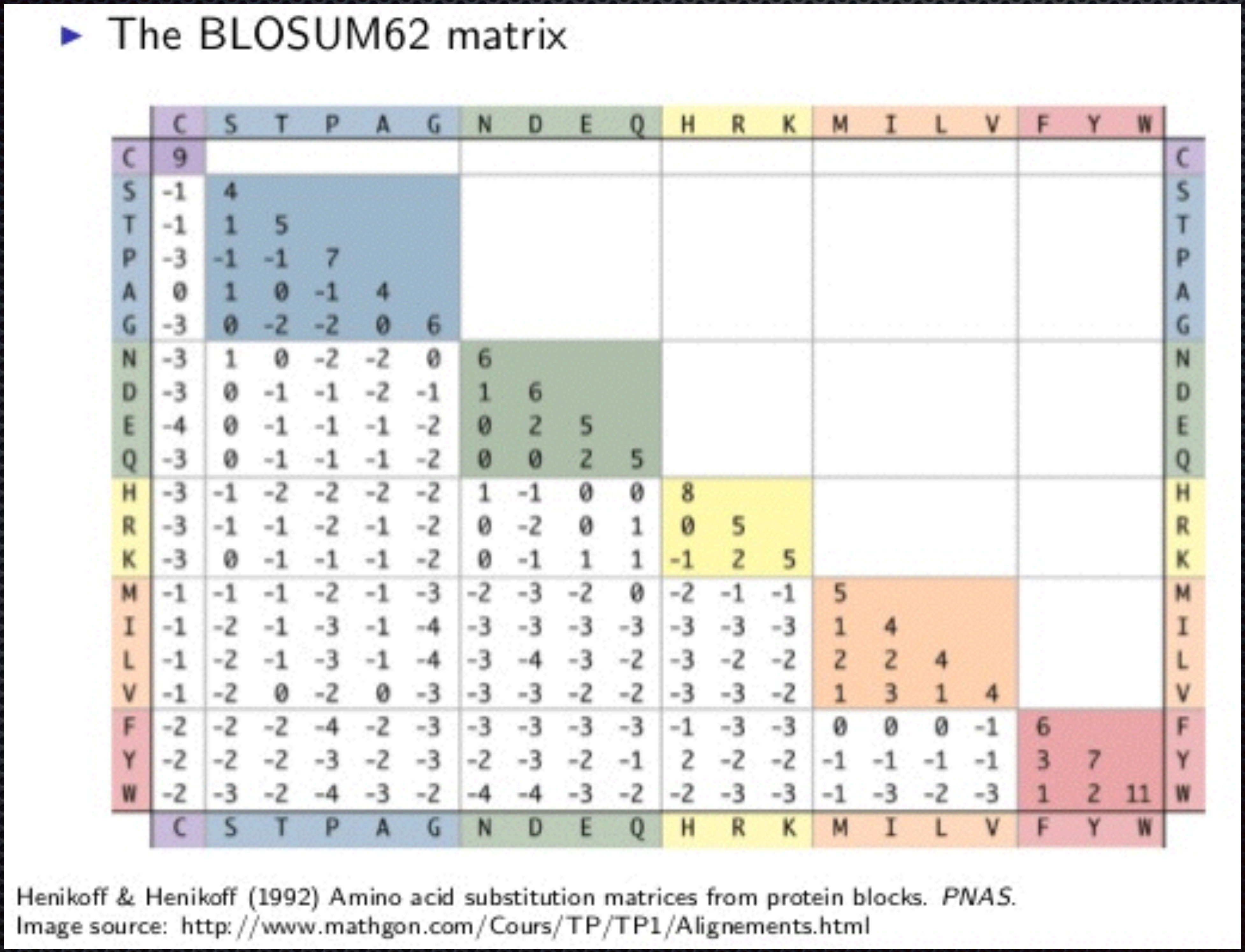


極性

芳香族

疎水性 www.russelllab.org

Similarity-scoring matrix



- The BLOSUM62 matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Henikoff & Henikoff (1992) Amino acid substitution matrices from protein blocks. *PNAS*.
Image source: http://www.mathgon.com/Cours/TP/TP1/Alignements.html

アミノ酸置換行列の最適化

類似配列検索

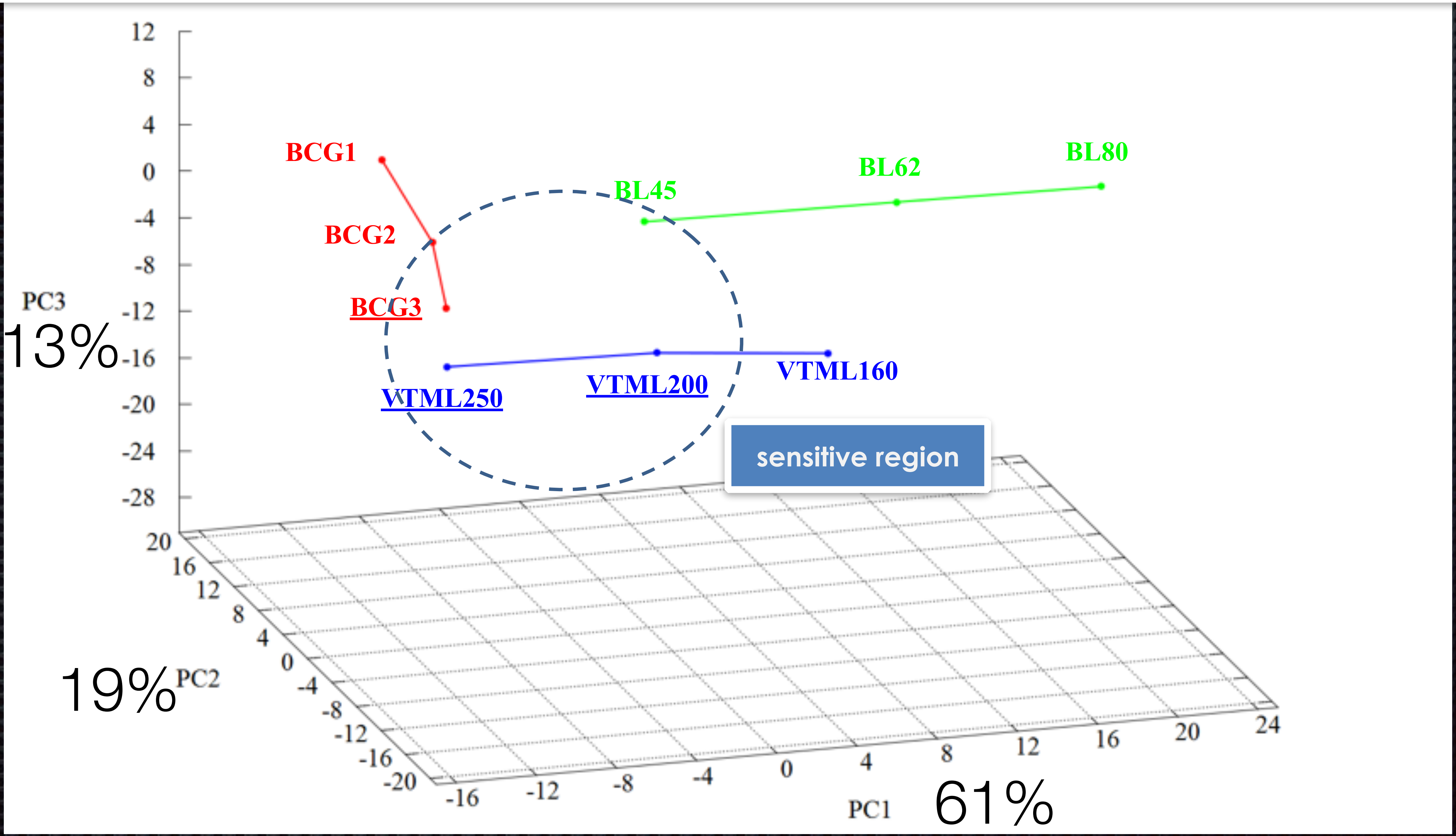
- 高速配列データベース検索手法
  - FASTA (faculty.virginia.edu/wrpearson/fasta/)
  - BLAST/PSI-BLAST (www.ncbi.nlm.nih.gov/BLAST/)

Scoring matrices

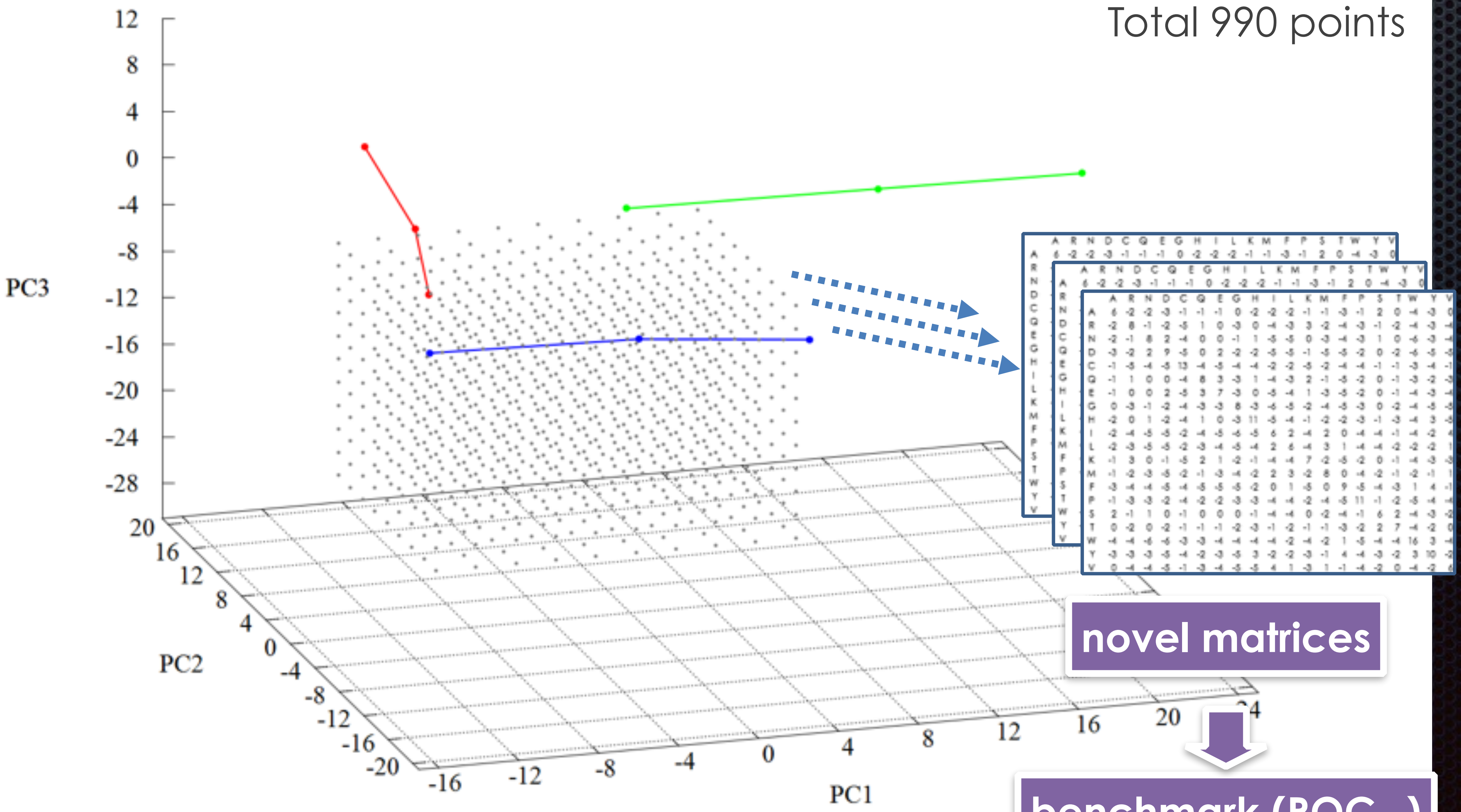
- **BLOSUM matrices (Henikoff & Henikoff, 1992)**
- **updated PAM matrices (Benner et al., 1994)**
- **VTML (Muller et al., 2002)**

Principal Component Analysis

PCA with existing 9 matrices. Obtained principal component score are plotted to PCA space (pc1-3 axes). A cumulative contribution ration of the 3 axes is about 93%.



Grid Search



Total 990 points

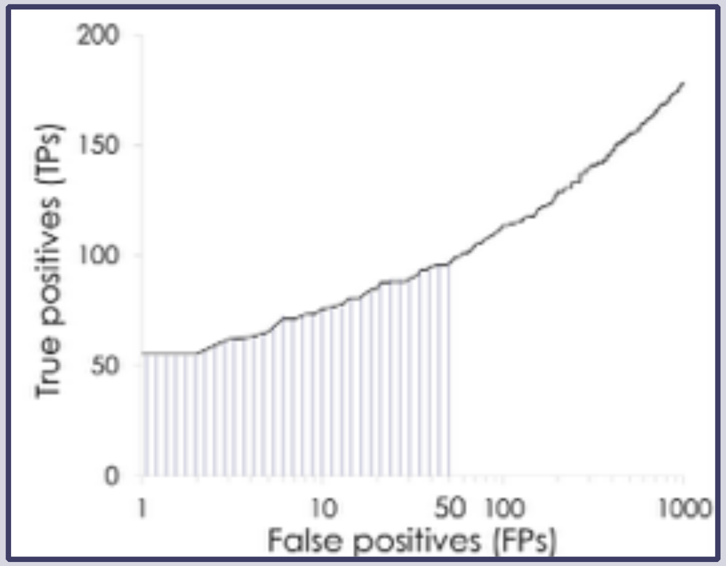
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	6	-2	-2	-3	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	2	0	-4	-3	0
R	-2	8	-1	-2	-5	1	0	-3	0	-4	-3	3	-2	-4	-3	-1	-2	-4	-3	-4
N	-2	-1	8	2	-4	0	0	-1	1	-5	-5	0	-3	-4	-3	1	0	-6	-3	-4
D	-3	-2	2	9	-5	0	2	-2	-2	-5	-5	-1	-5	-5	-2	0	-2	-6	-5	-5
C	-1	-5	-4	-5	13	-4	-5	-4	-4	-2	-2	-5	-2	-4	-4	-1	-1	-3	-4	-1
Q	-1	1	0	0	-4	8	3	-3	1	-4	-3	2	-1	-5	-2	0	-1	-3	-2	-3
E	-1	0	0	2	-5	3	7	-3	0	-5	-4	1	-3	-5	-2	0	-1	-4	-3	-4
G	0	-3	-1	-2	-4	-3	-3	8	-3	-6	-5	-2	-4	-5	-3	0	-2	-4	-5	-5
H	-2	0	1	-2	-4	1	0	-3	11	-5	-4	-1	-2	-2	-3	-1	-3	-4	3	-5
I	-2	-4	-5	-5	-2	-4	-5	-6	-5	6	2	-4	2	0	-4	-4	-1	-4	-2	4
L	-2	-3	-5	-5	-2	-3	-4	-5	-4	2	6	-4	3	1	-4	-4	-2	-2	-2	1
K	-1	3	0	-1	-5	2	1	-2	-1	-4	-4	7	-2	-5	-2	0	-1	-4	-3	-3
M	-1	-2	-3	-5	-2	-1	-3	-4	-2	2	3	-2	8	0	-4	-2	-1	-2	-1	1
F	-3	-4	-4	-5	-4	-5	-5	-5	-2	0	1	-5	0	9	-5	-4	-3	1	4	-1
P	-1	-3	-3	-2	-4	-2	-2	-3	-3	-4	-4	-2	-4	-5	11	-1	-2	-5	-4	-4
S	2	-1	1	0	-1	0	0	0	-1	-4	-4	0	-2	-4	-1	6	2	-4	-3	-2
T	0	-2	0	-2	-1	-1	-1	-2	-3	-1	-2	-1	-1	-3	-2	2	7	-4	-2	0
W	-4	-4	-6	-6	-3	-3	-4	-4	-4	-4	-2	-4	-2	1	-5	-4	-4	16	3	-4
Y	-3	-3	-3	-5	-4	-2	-3	-5	3	-2	-2	-3	-1	4	-4	-3	-2	3	10	-2
V	0	-4	-4	-5	-1	-3	-4	-5	-5	4	1	-3	1	-1	-4	-2	0	-4	-2	6

novel matrices

benchmark (ROC$_{50}$)

Method: benchmark

alignment algorithm: 方法:	SSEARCH (local aligner) 全対全検索
トレーニングセット:	**SCOP20** (ランダムに選択した3537配列)
テストセット:	**SCOP20** (残りの3537配列)
正誤の判定:	正解⇔SFの一致、不正解⇔Foldの不一致
検出感度の評価:	**ROC_{50}**



$$ROC_{50} = \frac{1}{50T}\sum_{i}^{50} t_i$$

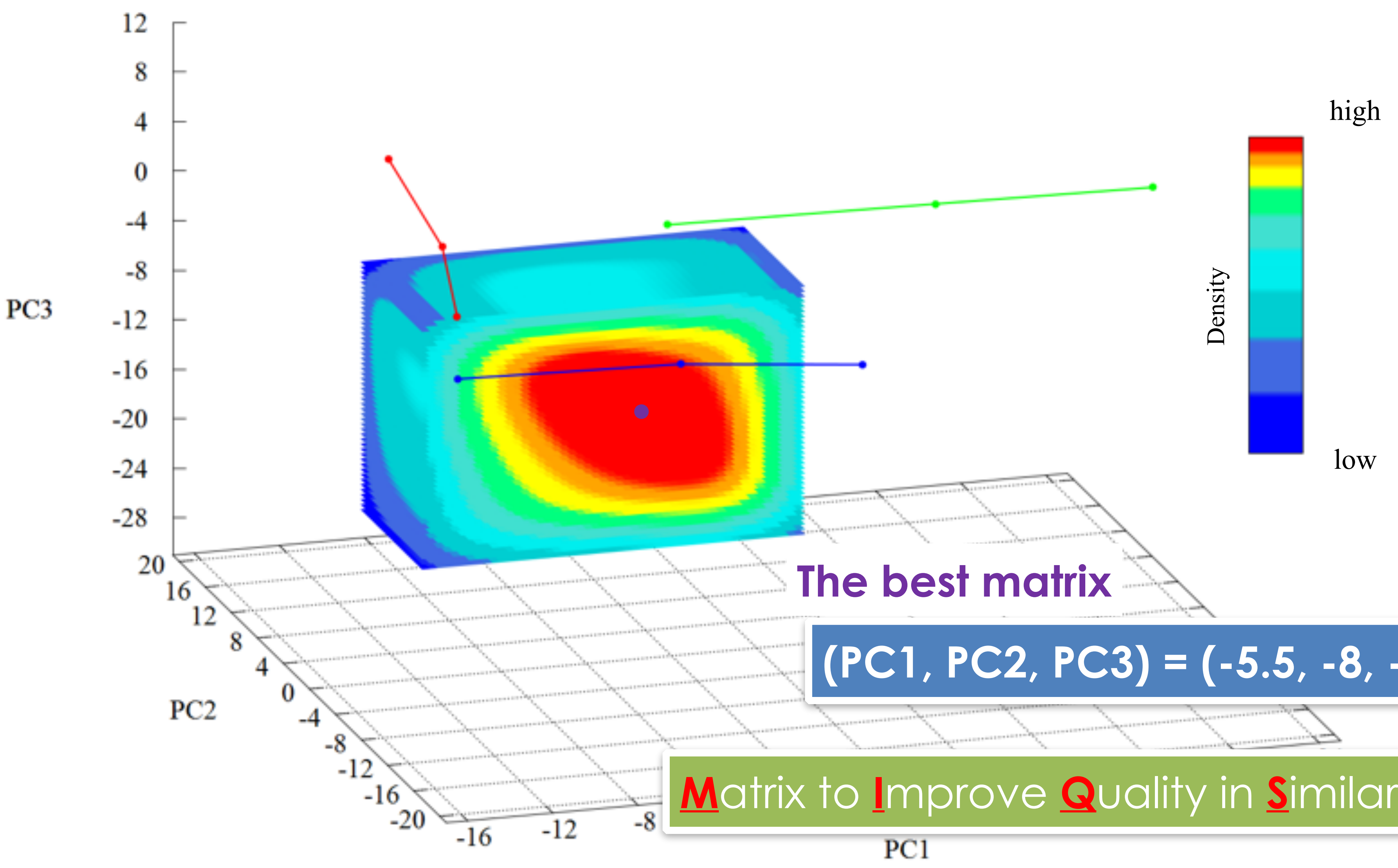
T: 全正解数

t_i: FPがi番目までのTP数

ギャップペナルティー:	開始 [-13, -9]、拡張 [-2, -1] (1刻み)

Kernel Density Estimation

Confined sensitive region

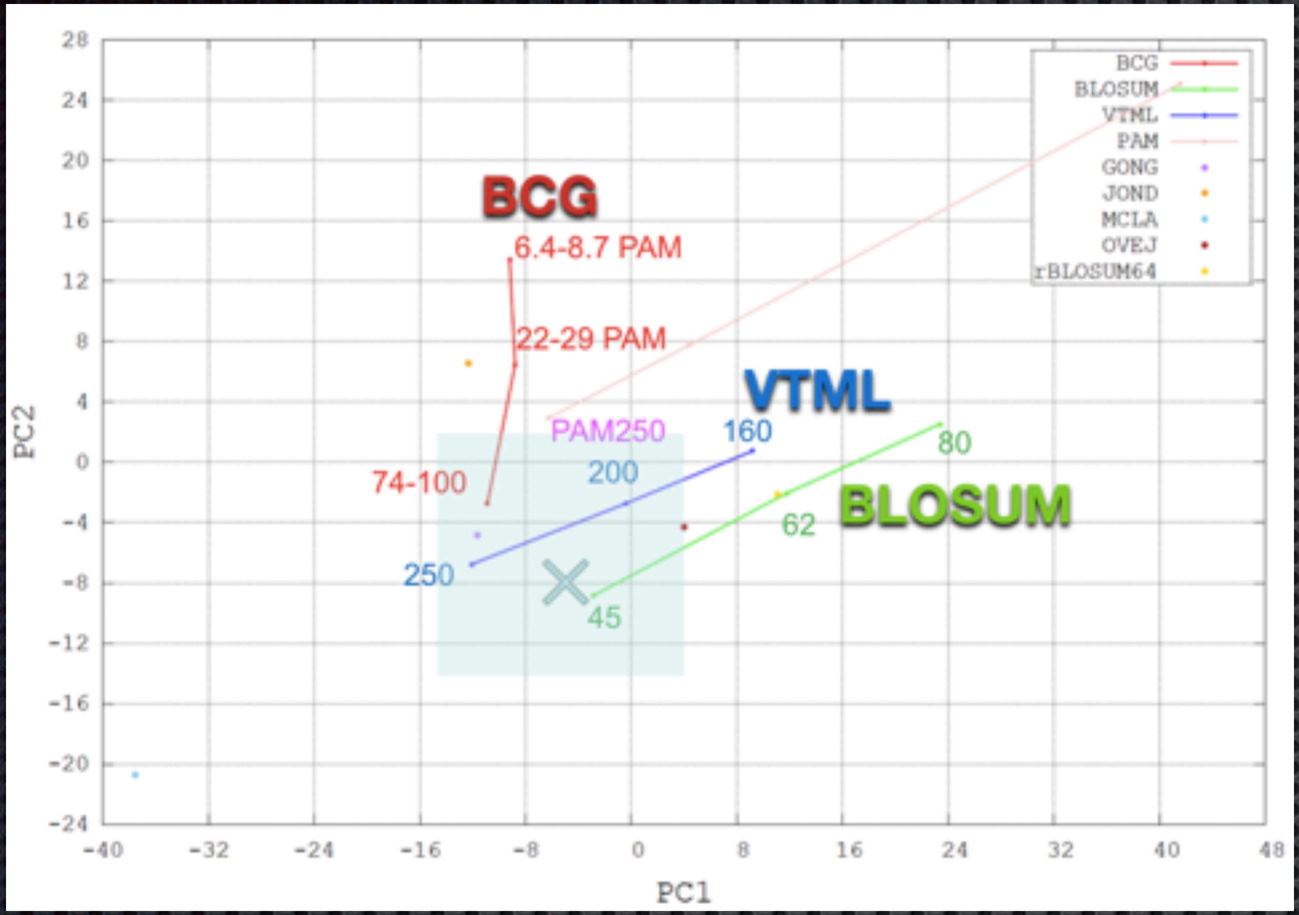


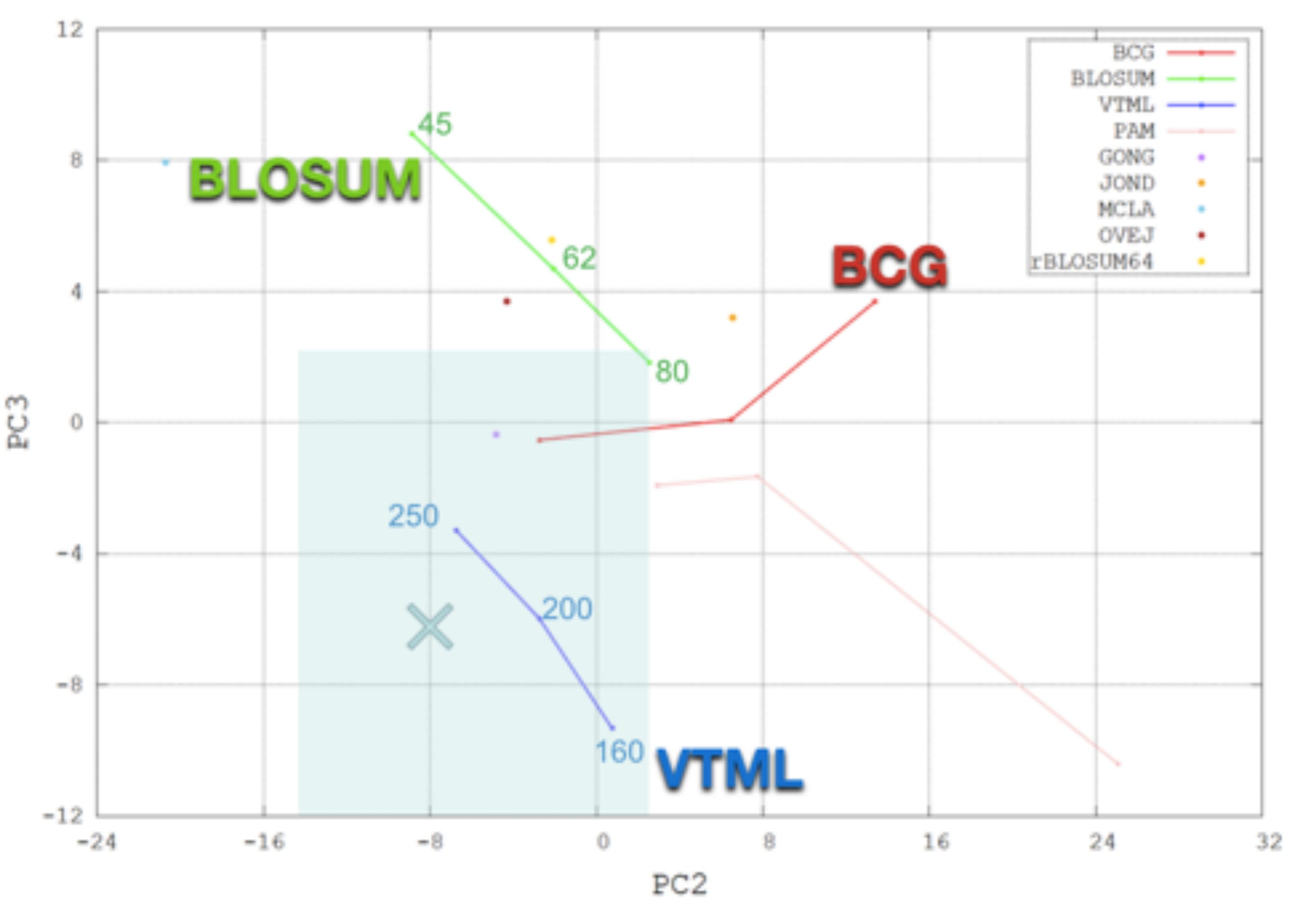
The best matrix

(PC1, PC2, PC3) = (-5.5, -8, -6.5)

Matrix to Improve Quality in Similarity search

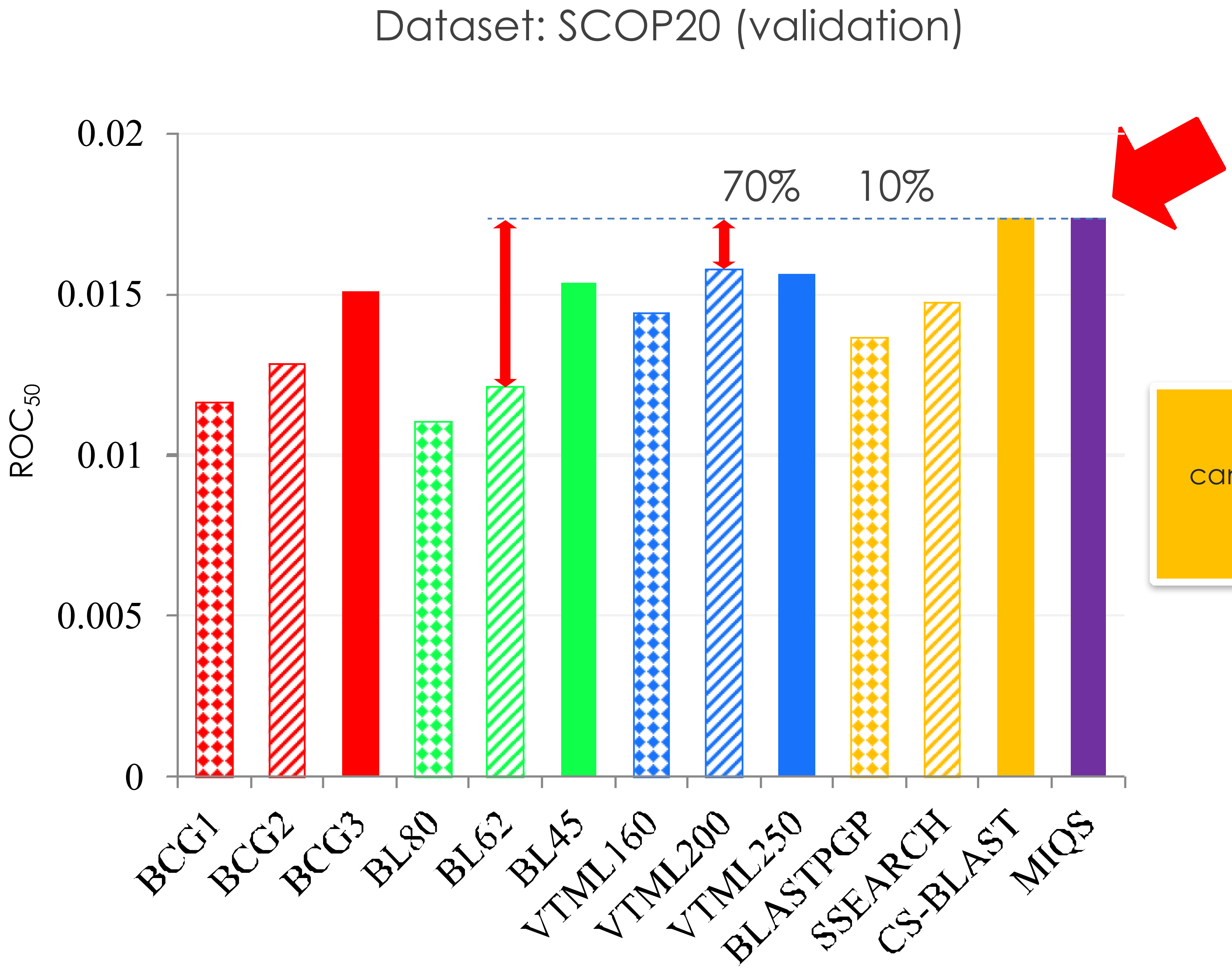
Yamada K, Tomii K (2014) *Bioinformatics* **30**:317–325.


BCG
6.4-8.7 PAM
22-29 PAM
74-100
PAM250
VTML
160
200
250
BLOSUM
80
62
45
BCG
BLOSUM
VTML
PAM
GONG
JOND
MCLA
OVEJ
rBLOSUM64
PC1
PC2
28
24
20
16
12
8
4
0
-4
-8
-12
-16
-20
-24
-40
-32
-24
-16
-8
0
8
16
24
32
40
48


BLOSUM
45
62
80
BCG
250
200
160
VTML
BCG
BLOSUM
VTML
PAM
GONG
JOND
MCLA
OVEJ
rBLOSUM64
PC2
PC3
12
8
4
0
-4
-8
-12
-24
-16
-8
0
8
16
24
32

Results

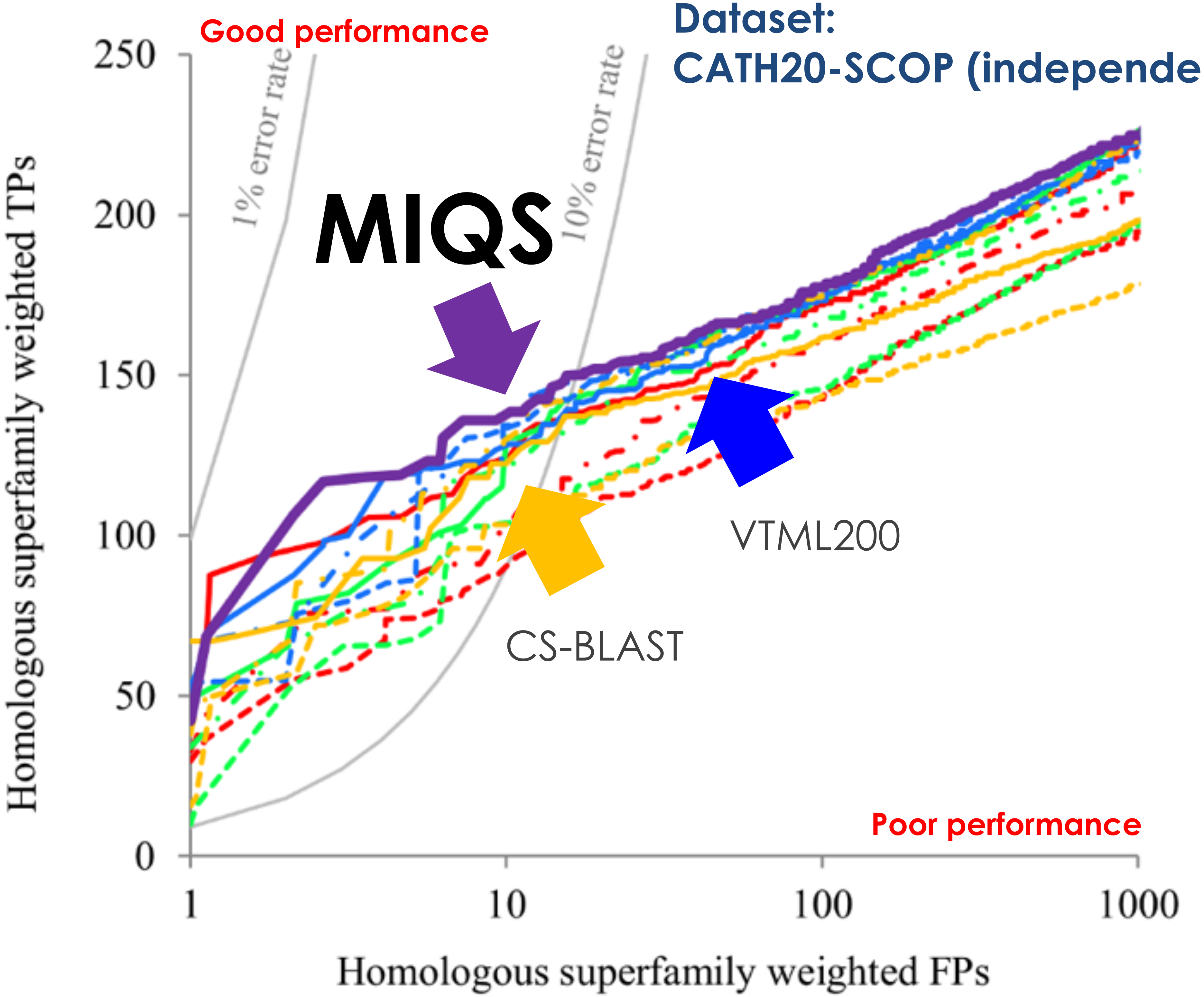
Validation dataset (SCOP20)



CS-BLAST
can search without any matrices, high performance

Test dataset (CATH20-SCOP)

Dataset:
CATH20-SCOP (independent data)



Good performance

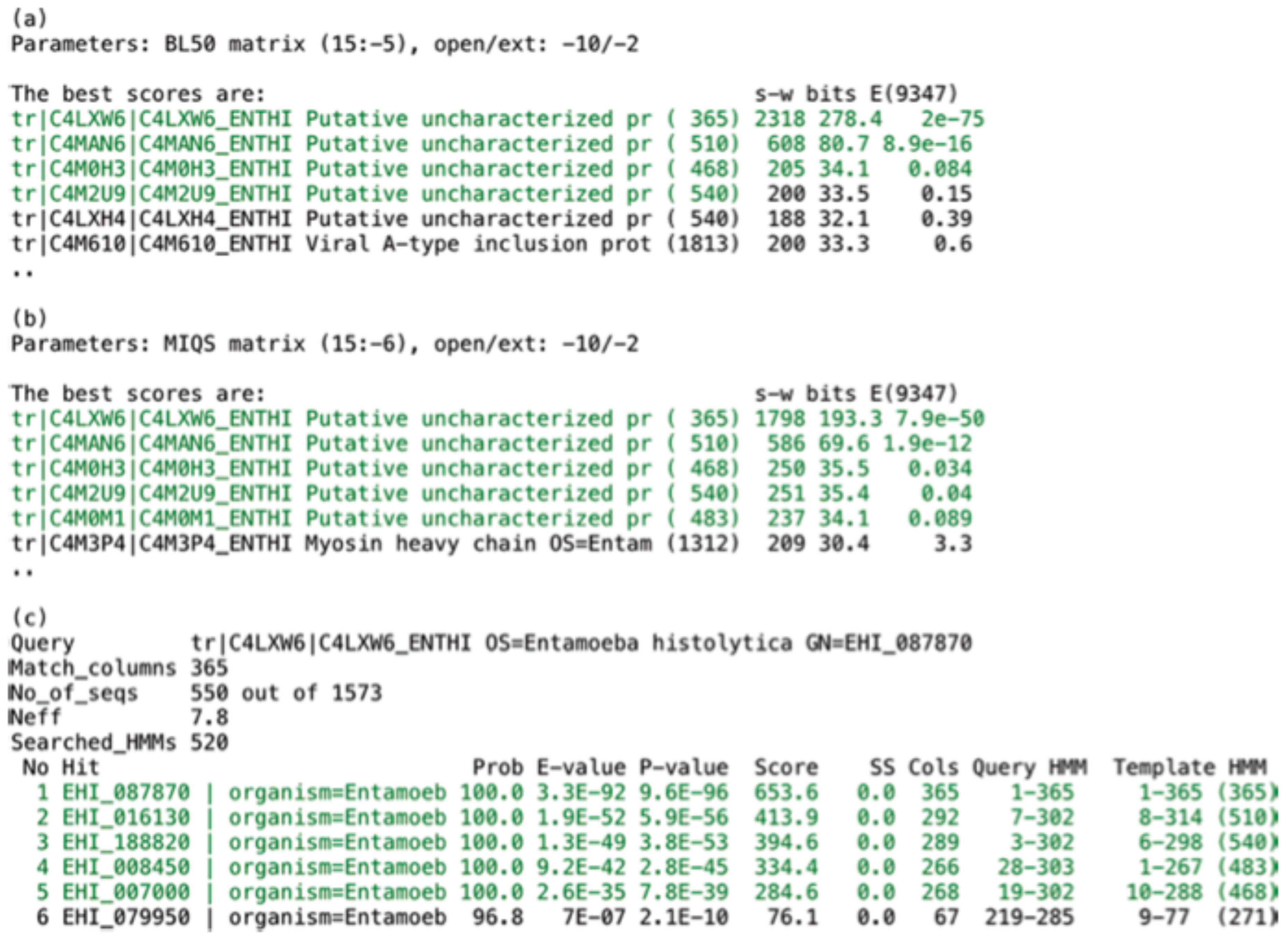
MIQS

CS-BLAST

VTML200

Poor performance

Yamada K, Tomii K (2014) *Bioinformatics* **30**:317–325.



```
(a)
Parameters: BL50 matrix (15:-5), open/ext: -10/-2

The best scores are:                                          s-w bits E(9347)
tr|C4LXW6|C4LXW6_ENTHI Putative uncharacterized pr ( 365) 2318 278.4   2e-75
tr|C4MAN6|C4MAN6_ENTHI Putative uncharacterized pr ( 510)  608 80.7 8.9e-16
tr|C4M0H3|C4M0H3_ENTHI Putative uncharacterized pr ( 468)  205 34.1   0.084
tr|C4M2U9|C4M2U9_ENTHI Putative uncharacterized pr ( 540)  200 33.5    0.15
tr|C4LXH4|C4LXH4_ENTHI Putative uncharacterized pr ( 540)  188 32.1    0.39
tr|C4M610|C4M610_ENTHI Viral A-type inclusion prot (1813)  200 33.3     0.6
..

(b)
Parameters: MIQS matrix (15:-6), open/ext: -10/-2

The best scores are:                                          s-w bits E(9347)
tr|C4LXW6|C4LXW6_ENTHI Putative uncharacterized pr ( 365) 1798 193.3 7.9e-50
tr|C4MAN6|C4MAN6_ENTHI Putative uncharacterized pr ( 510)  586 69.6 1.9e-12
tr|C4M0H3|C4M0H3_ENTHI Putative uncharacterized pr ( 468)  250 35.5   0.034
tr|C4M2U9|C4M2U9_ENTHI Putative uncharacterized pr ( 540)  251 35.4    0.04
tr|C4M0M1|C4M0M1_ENTHI Putative uncharacterized pr ( 483)  237 34.1   0.089
tr|C4M3P4|C4M3P4_ENTHI Myosin heavy chain OS=Entam (1312)  209 30.4     3.3
..

(c)
Query          tr|C4LXW6|C4LXW6_ENTHI OS=Entamoeba histolytica GN=EHI_087870
Match_columns 365
No_of_seqs    550 out of 1573
Neff          7.8
Searched_HMMs 520
 No Hit                             Prob E-value P-value  Score    SS Cols Query HMM  Template HMM
  1 EHI_087870 | organism=Entamoeb 100.0 3.3E-92 9.6E-96  653.6   0.0  365    1-365     1-365 (365)
  2 EHI_016130 | organism=Entamoeb 100.0 1.9E-52 5.9E-56  413.9   0.0  292    7-302     8-314 (510)
  3 EHI_188820 | organism=Entamoeb 100.0 1.3E-49 3.8E-53  394.6   0.0  289    3-302     6-298 (540)
  4 EHI_008450 | organism=Entamoeb 100.0 9.2E-42 2.8E-45  334.4   0.0  266   28-303     1-267 (483)
  5 EHI_007000 | organism=Entamoeb 100.0 2.6E-35 7.8E-39  284.6   0.0  268   19-302    10-288 (468)
  6 EHI_079950 | organism=Entamoeb  96.8   7E-07 2.1E-10   76.1   0.0   67  219-285      9-77  (271)
```

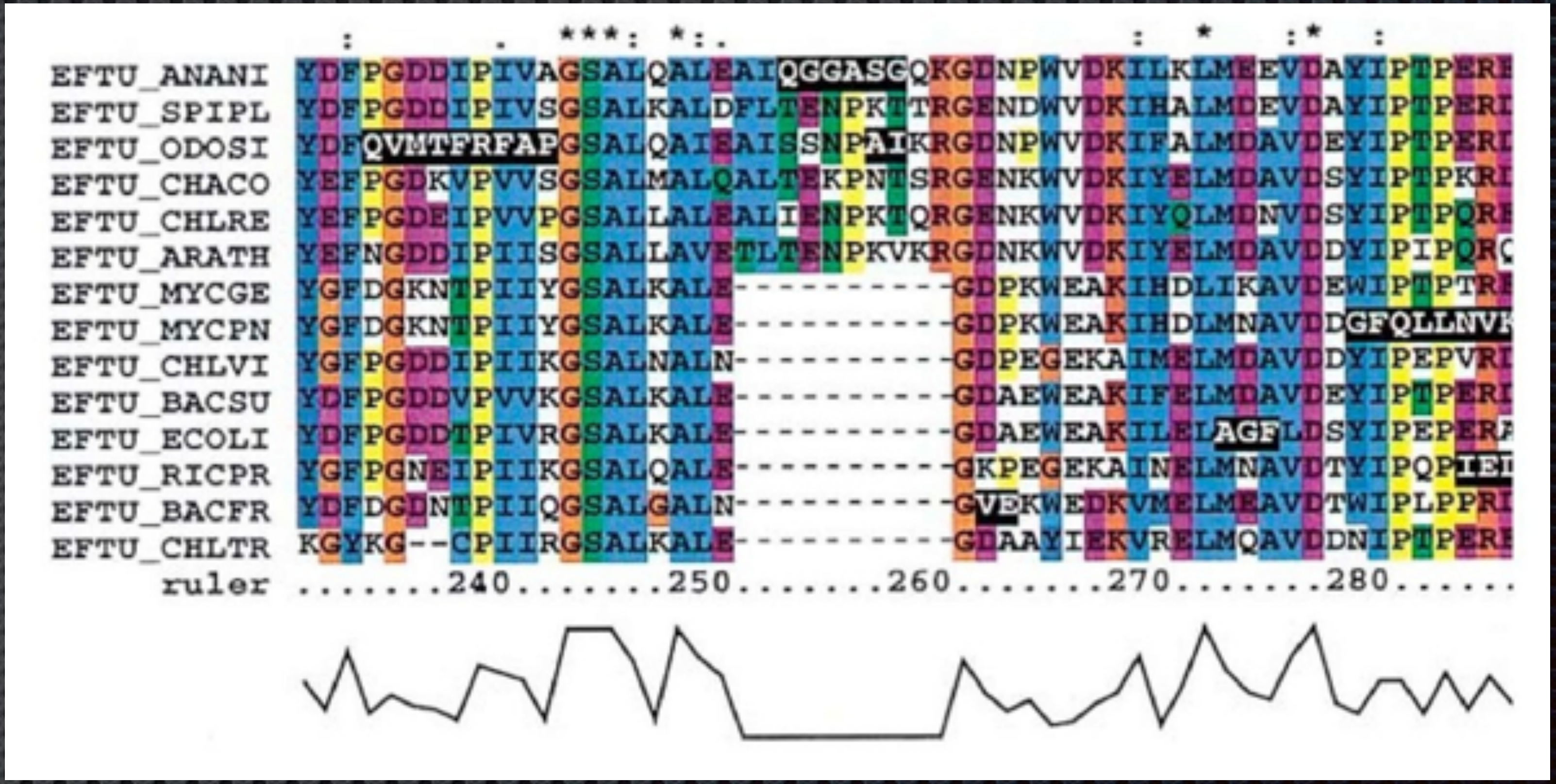
Fig. 3 Similarity search results of EHI_087870 against the *Entamoeba histolytica* proteome. Proteins detected by the SSEARCH program with the default setting, i.e., with BLOSUM50 (**a**) and with MIQS (**b**), are shown. (**c**) Proteins detected using HHblits are shown. Putative IMD/I-BAR domain-containing proteins in *E. histolytica* are shown in *green*

Oliviero Carugo and Frank Eisenhaber (eds.), *Data Mining Techniques for the Life Sciences*, Methods in Molecular Biology, vol. 1415, DOI 10.1007/978-1-4939-3572-7_11, © Springer Science+Business Media New York 2016

Multiple Sequence Alignment (MSA)

Multiple sequence alignment

- 機能推定
- 立体構造推定
- 機能部位推定
- 分子系統解析



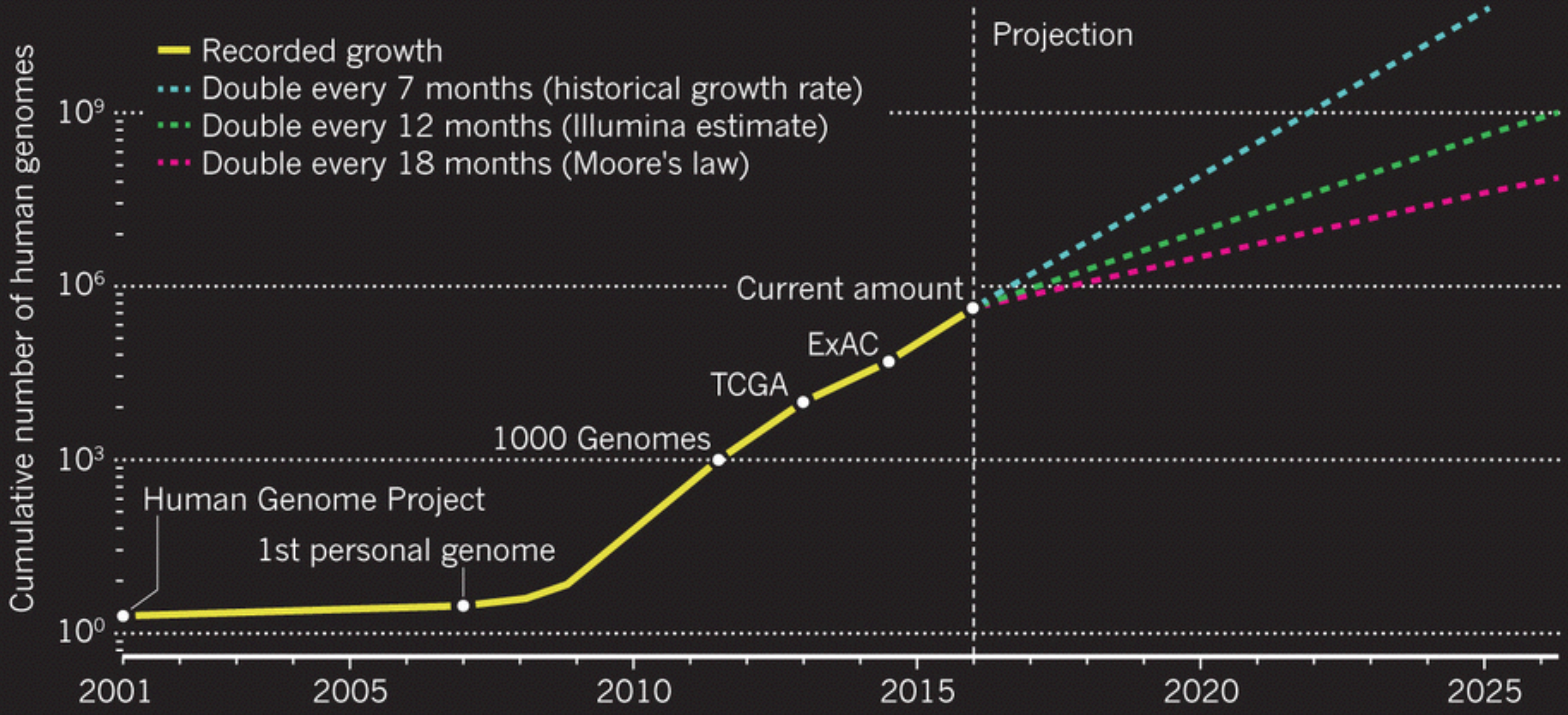
```
                    :     .   ***:  *:.                     :   *    :*  :
EFTU_ANANI  YDFPGDDIPIVAGSALQALEAIQGGASGQKGDNPWVDKILKLMEEVDAYIPTPERE
EFTU_SPIPL  YDFPGDDIPIVSGSALKALDFLTENPKTTRGENDWVDKIHALMDEVDAYIPTPERI
EFTU_ODOSI  YDFQVMTFRFAPGSALQAIEAISSNPAIKRGDNPWVDKIFALMDAVDEYIPTPERI
EFTU_CHACO  YEFPGDKVPVVSGSALMALQALTEKPNTSRGENKWVDKIYELMDAVDSYIPTPKRI
EFTU_CHLRE  YEFPGDEIPVVPGSALLALEALIENPKTQRGENKWVDKIYQLMDNVDSYIPTPQRE
EFTU_ARATH  YEFNGDDIPIISGSALLAVETLTENPKVKRGDNKWVDKIYELMDAVDDYIPIPQRQ
EFTU_MYCGE  YGFDGKNTPIIYGSALKALE----------GDPKWEAKIHDLIKAVDEWIPTPTRE
EFTU_MYCPN  YGFDGKNTPIIYGSALKALE----------GDPKWEAKIHDLMNAVDDGFQLLNVK
EFTU_CHLVI  YGFPGDDIPIIKGSALNALN----------GDPEGEKAIMELMDAVDDYIPEPVRI
EFTU_BACSU  YDFPGDDVPVVKGSALKALE----------GDAEWEAKIFELMDAVDEYIPTPERI
EFTU_ECOLI  YDFPGDDTPIVRGSALKALE----------GDAEWEAKILELAGFLDSYIPEPERA
EFTU_RICPR  YGFPGNEIPIIKGSALQALE----------GKPEGEKAINELMNAVDTYIPQPIED
EFTU_BACFR  YDFDGDNTPIIQGSALGALN----------GVEKWEDKVMELMEAVDTWIPLPPRI
EFTU_CHLTR  KGYKG--CPIIRGSALKALE----------GDAAYIEKVRELMQAVDDNIPTPERE
     ruler  .......240.......250.......260.......270.......280......
```

http://what-when-how.com/molecular-biology/aligning-sequences-molecular-biology/

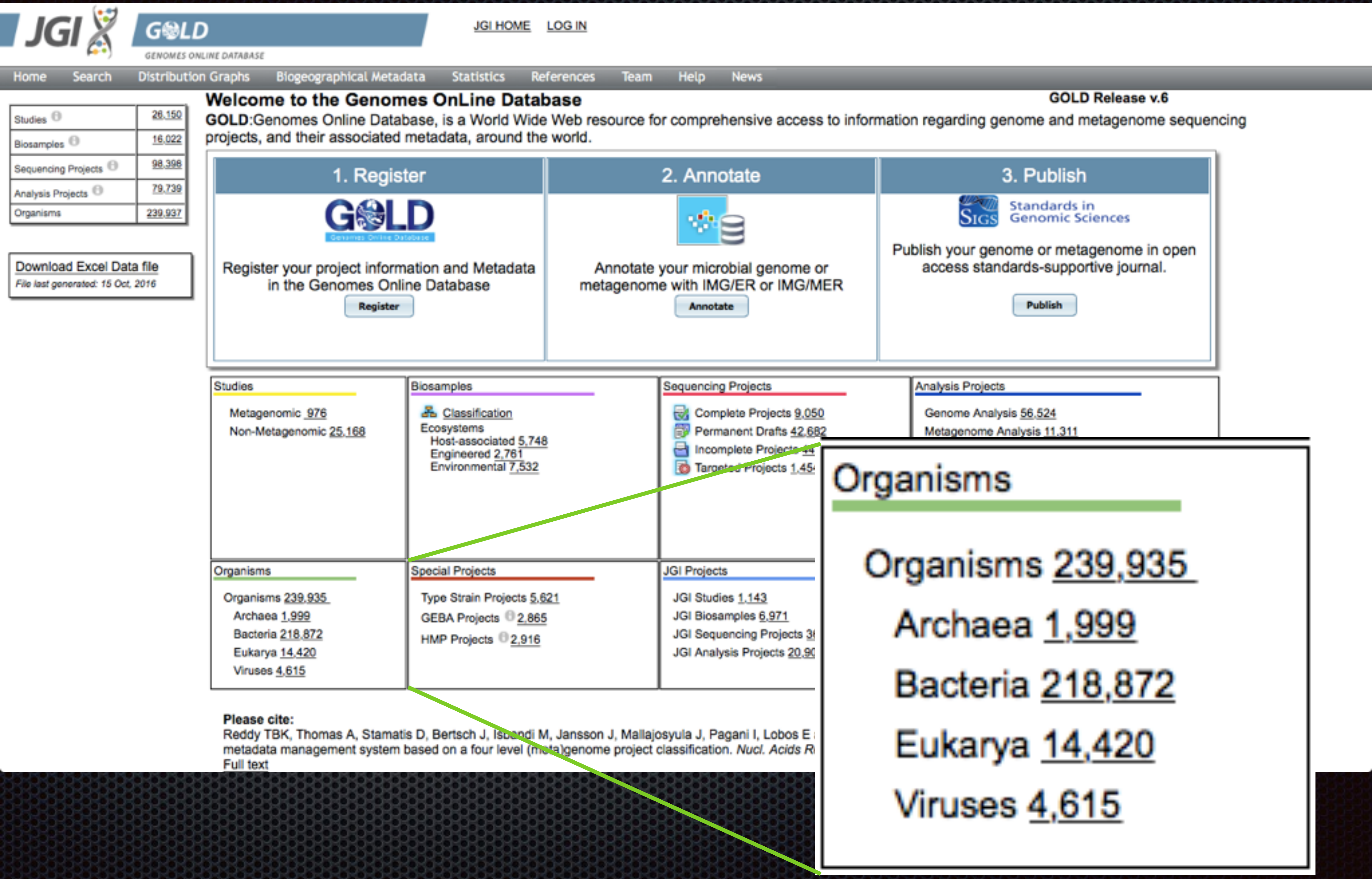
Big data: The power of petabytes

DNA SEQUENCING SOARS

Human genomes are being sequenced at an ever-increasing rate. The 1000 Genomes Project has aggregated hundreds of genomes; The Cancer Genome Atlas (TGCA) has gathered several thousand; and the Exome Aggregation Consortium (ExAC) has sequenced more than 60,000 exomes. Dotted lines show three possible future growth curves.



Nature **527**, S2–S4 (05 November 2015)



Studies	26,150
Biosamples	16,022
Sequencing Projects	98,398
Analysis Projects	79,739
Organisms	239,937

Download Excel Data file
File last generated: 15 Oct, 2016

Welcome to the Genomes OnLine Database

GOLD Release v.6

GOLD:Genomes Online Database, is a World Wide Web resource for comprehensive access to information regarding genome and metagenome sequencing projects, and their associated metadata, around the world.

1. Register

Register your project information and Metadata in the Genomes Online Database

2. Annotate

Annotate your microbial genome or metagenome with IMG/ER or IMG/MER

3. Publish

Publish your genome or metagenome in open access standards-supportive journal.

Studies
Metagenomic 976
Non-Metagenomic 25,168

Biosamples
Classification
Ecosystems
Host-associated 5,748
Engineered 2,761
Environmental 7,532

Sequencing Projects
Complete Projects 9,050
Permanent Drafts 42,682
Incomplete Projects 44
Targeted Projects 1,45

Analysis Projects
Genome Analysis 56,524
Metagenome Analysis 11,311

Organisms
Organisms 239,935
Archaea 1,999
Bacteria 218,872
Eukarya 14,420
Viruses 4,615

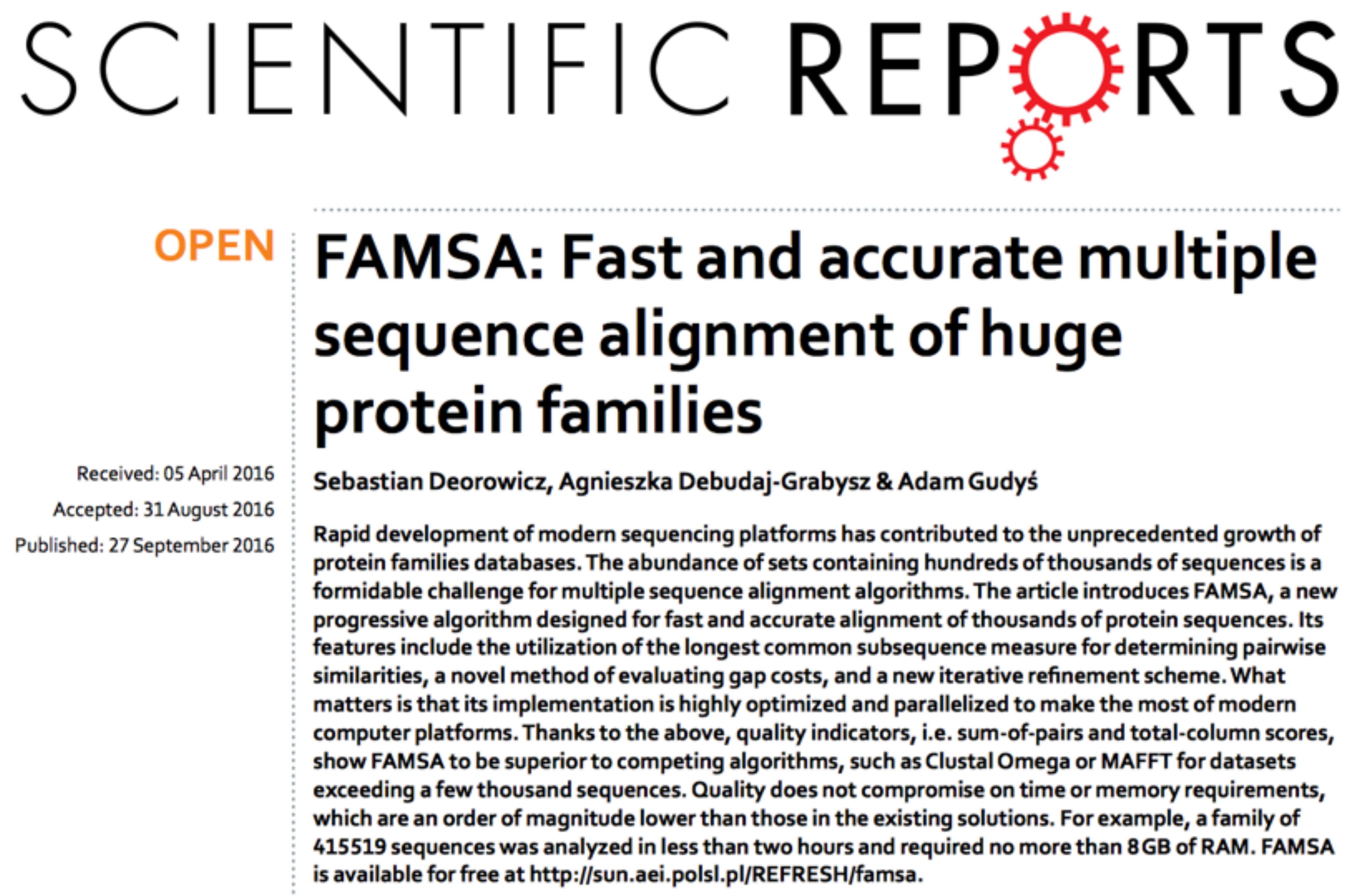
Special Projects
Type Strain Projects 5,621
GEBA Projects 2,865
HMP Projects 2,916

JGI Projects
JGI Studies 1,143
JGI Biosamples 6,971
JGI Sequencing Projects 3
JGI Analysis Projects 20,9

Please cite:
Reddy TBK, Thomas A, Stamatis D, Bertsch J, Isbandi M, Jansson J, Mallajosyula J, Pagani I, Lobos E
metadata management system based on a four level (meta)genome project classification. *Nucl. Acids R*
Full text

Organisms
Organisms 239,935
Archaea 1,999
Bacteria 218,872
Eukarya 14,420
Viruses 4,615

MIQS used in MSA



SCIENTIFIC REPORTS

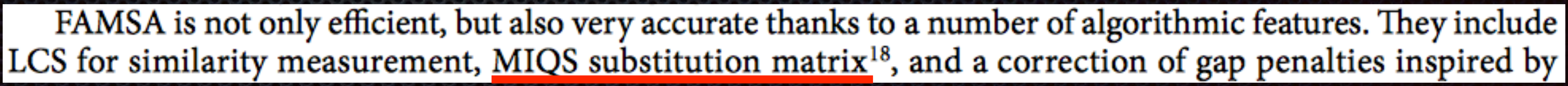
OPEN

FAMSA: Fast and accurate multiple sequence alignment of huge protein families

Received: 05 April 2016
Accepted: 31 August 2016
Published: 27 September 2016

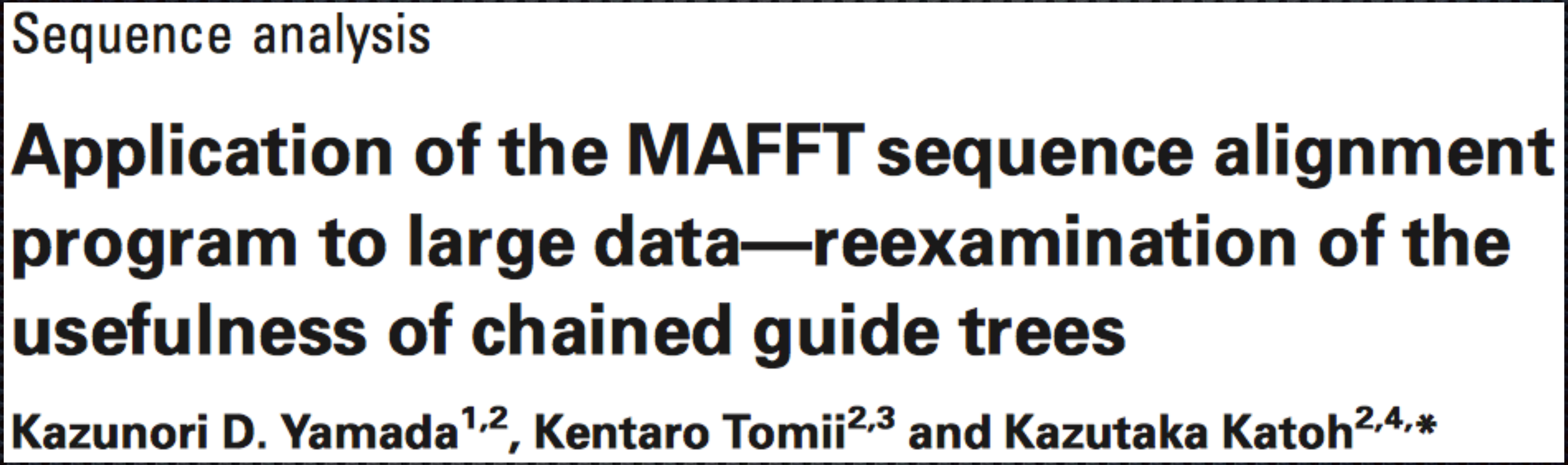
Sebastian Deorowicz, Agnieszka Debudaj-Grabysz & Adam Gudyś

Rapid development of modern sequencing platforms has contributed to the unprecedented growth of protein families databases. The abundance of sets containing hundreds of thousands of sequences is a formidable challenge for multiple sequence alignment algorithms. The article introduces FAMSA, a new progressive algorithm designed for fast and accurate alignment of thousands of protein sequences. Its features include the utilization of the longest common subsequence measure for determining pairwise similarities, a novel method of evaluating gap costs, and a new iterative refinement scheme. What matters is that its implementation is highly optimized and parallelized to make the most of modern computer platforms. Thanks to the above, quality indicators, i.e. sum-of-pairs and total-column scores, show FAMSA to be superior to competing algorithms, such as Clustal Omega or MAFFT for datasets exceeding a few thousand sequences. Quality does not compromise on time or memory requirements, which are an order of magnitude lower than those in the existing solutions. For example, a family of 415519 sequences was analyzed in less than two hours and required no more than 8GB of RAM. FAMSA is available for free at http://sun.aei.polsl.pl/REFRESH/famsa.



FAMSA is not only efficient, but also very accurate thanks to a number of algorithmic features. They include LCS for similarity measurement, MIQS substitution matrix[18], and a correction of gap penalties inspired by

Large multiple sequence alignments (MSAs)



Sequence analysis

Application of the MAFFT sequence alignment program to large data—reexamination of the usefulness of chained guide trees

Kazunori D. Yamada[1,2], Kentaro Tomii[2,3] and Kazutaka Katoh[2,4,*]

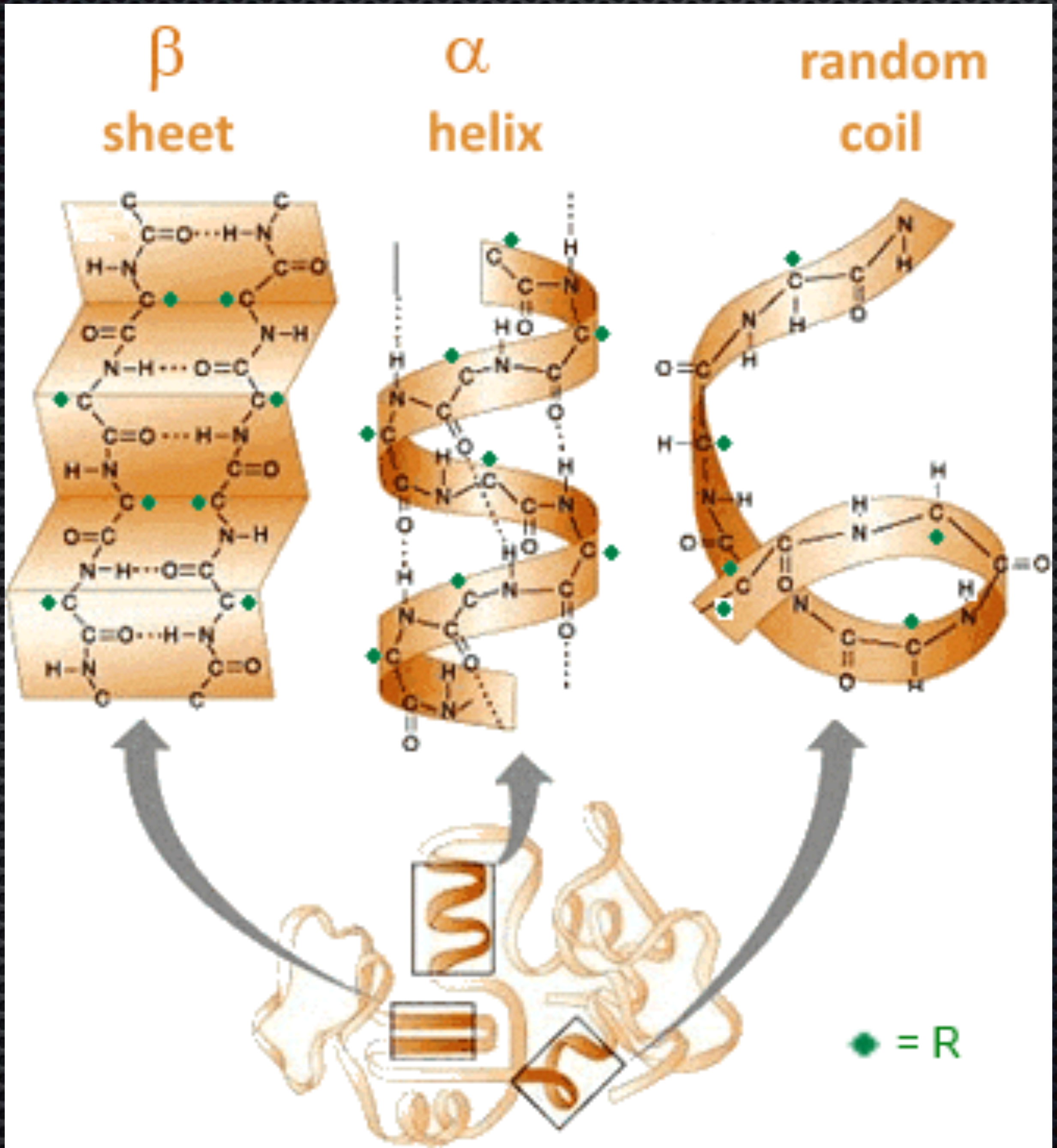
Large (N > 10,000), where N is the number of sequences in an MSA

Bioinformatics (2016)

Artificial intelligence for Bioinformatics

二次構造予測を例として

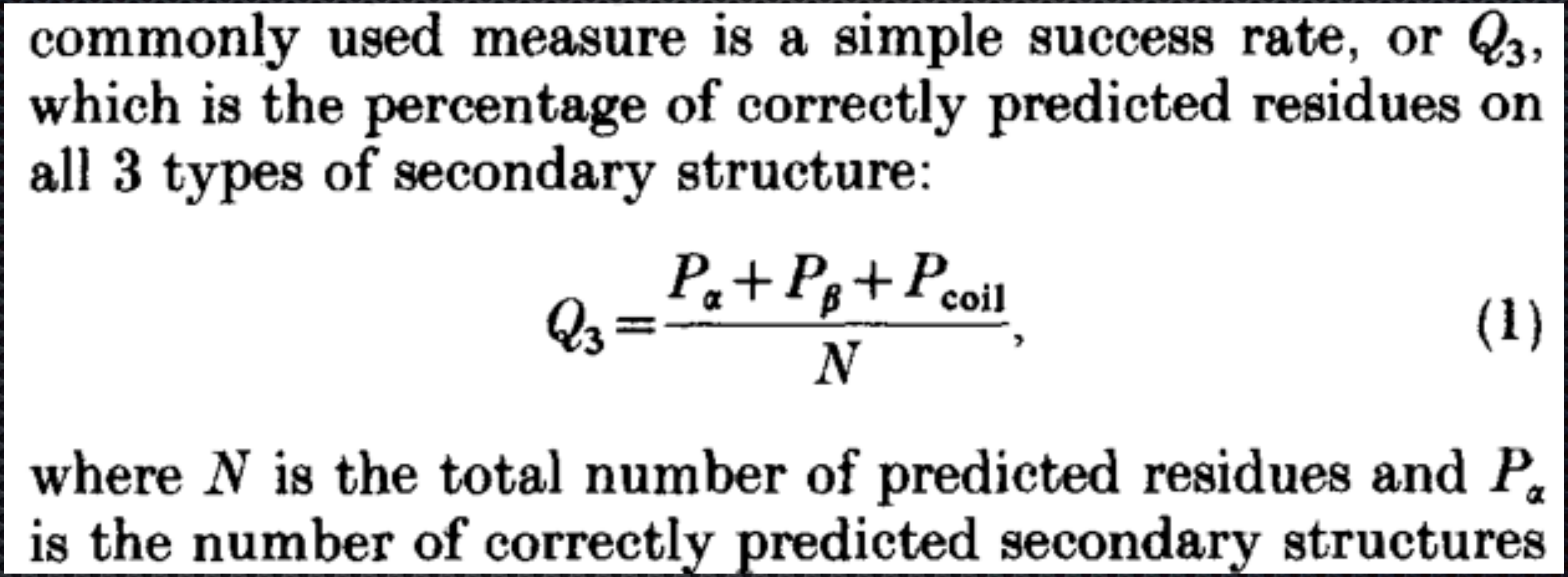
Secondary Structure Prediction



eprotstruct2.png @ chim.lu

Secondary Structure Prediction

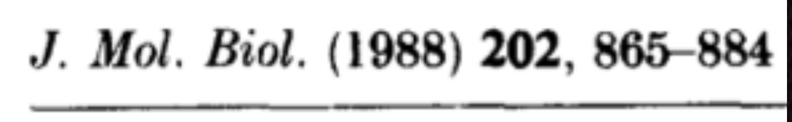
予測精度の指標(の一つ) Q_3



commonly used measure is a simple success rate, or Q_3, which is the percentage of correctly predicted residues on all 3 types of secondary structure:

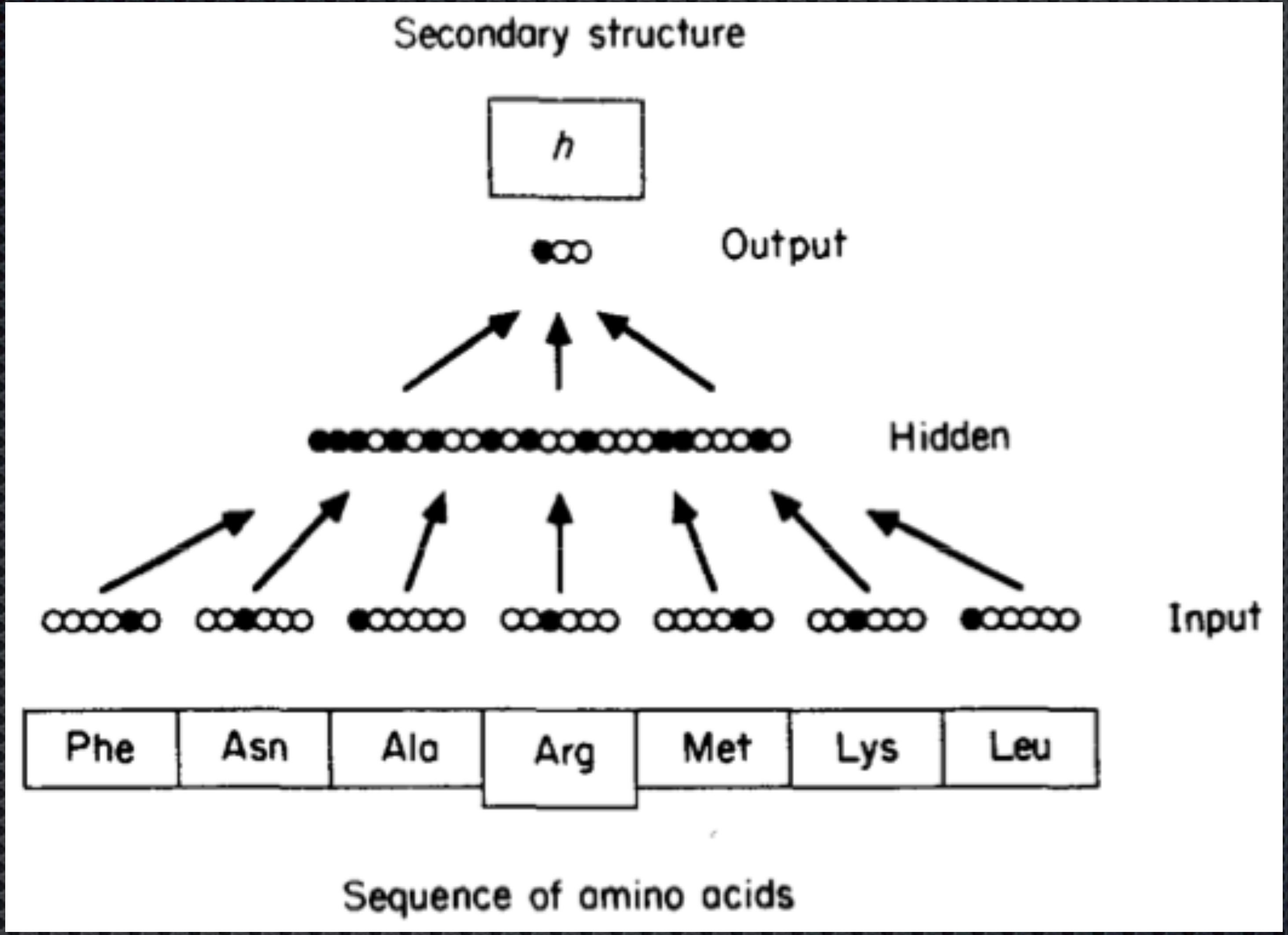
$$Q_3 = \frac{P_\alpha + P_\beta + P_{\mathrm{coil}}}{N}, \qquad (1)$$

where N is the total number of predicted residues and P_α is the number of correctly predicted secondary structures

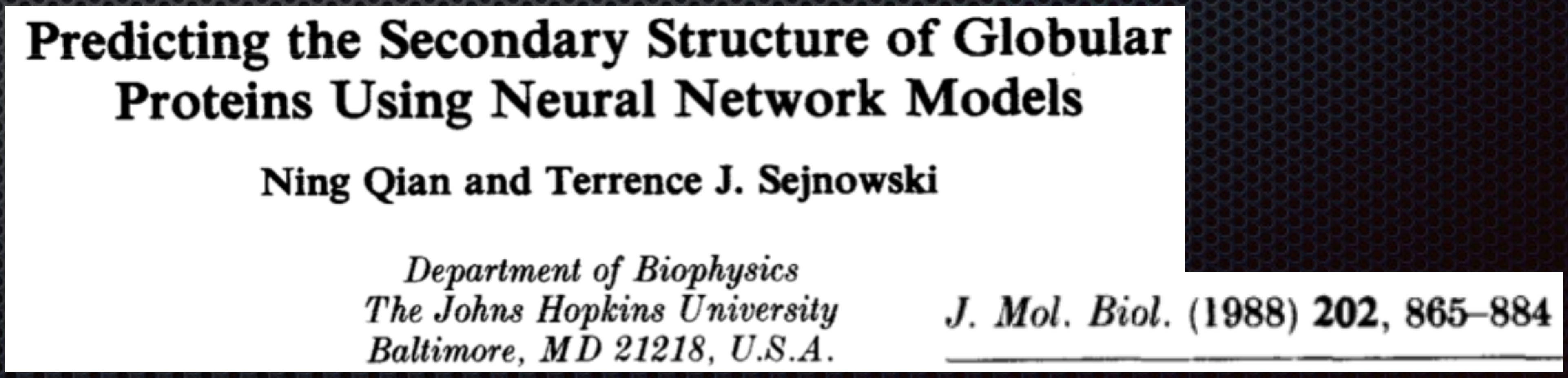


J. Mol. Biol. (1988) **202**, 865–884

28 years ago



$Q_3 = 64.3\%$



Predicting the Secondary Structure of Globular Proteins Using Neural Network Models

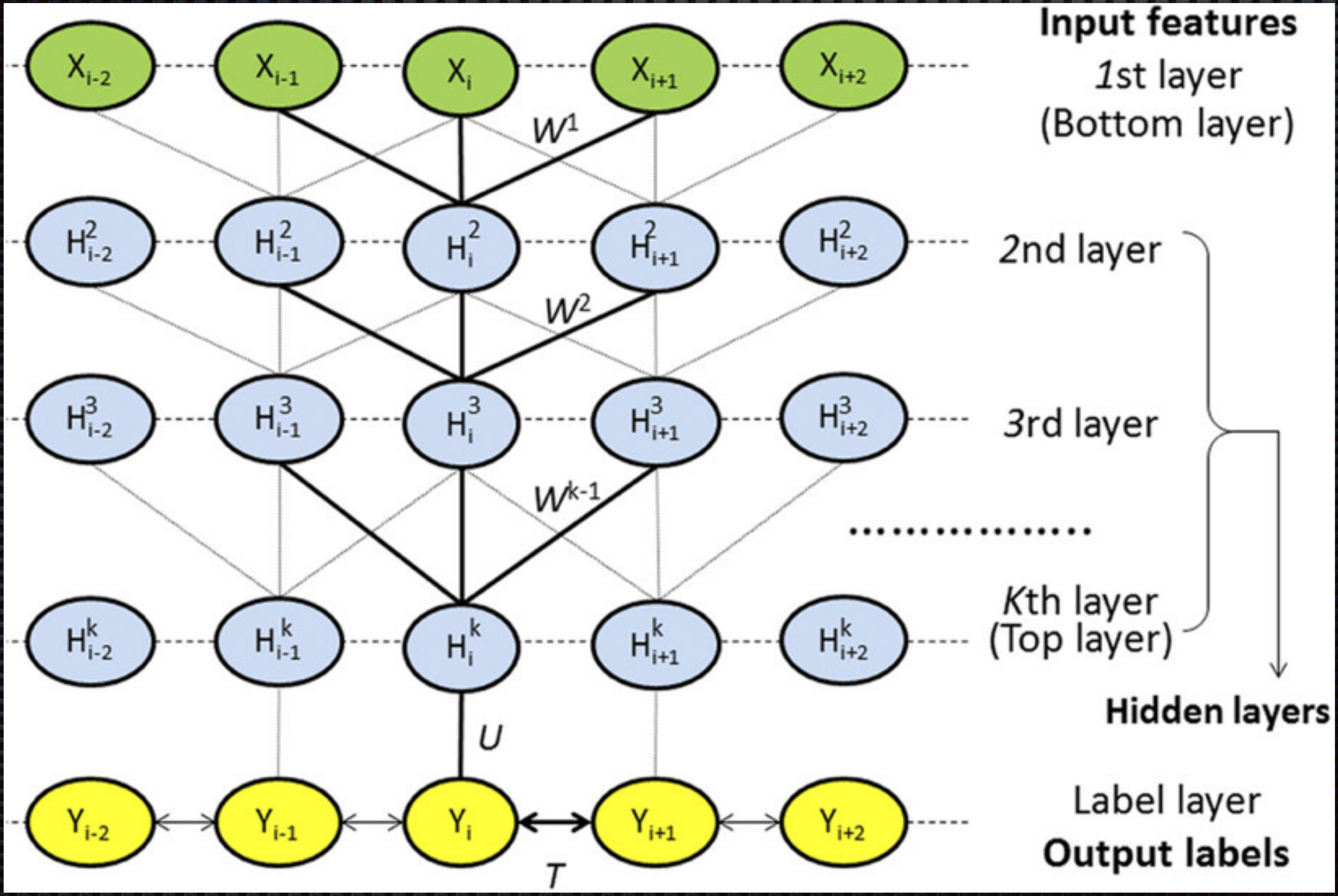
Ning Qian and Terrence J. Sejnowski

Department of Biophysics
The Johns Hopkins University
Baltimore, MD 21218, U.S.A.

J. Mol. Biol. (1988) **202**, 865–884

DeepCNF can obtain **~84%** Q_3 accuracy

and now (2016) ...



Sci. Rep. **6**, Article #: 18962 (2016)

Summary (新時代の計算生物学)

- 近年の配列データの著しい増大につれ、より高速、より大量、より正確な計算法が求められている。
  - 配列アラインメント
    - アミノ酸置換行列
- 計算生物学の分野でもAIの利用が加速中
  - 二次構造予測

“Thank you for your attention!”

Tomii Lab (http://cbrc3.cbrc.jp/~tomii/lab/)

謝辞

- 研究開発施設共用等促進費補助金（創薬等ライフサイエンス研究支援基盤事業）「創薬等支援技術基盤プラットフォーム事業」

創薬等支援技術基盤プラットフォーム
Platform for Drug Discovery, Informatics, and Structural Life Science