Macromolecules
Protein bioinformatics and modelling

Csaba Magyar, Institute of Enzymology,
Research Centre for Natural Sciences
2019 November

Bioinformatics

Interdisciplinary field, computational methods dealing with biological data

Experimental data handling

- drug trials: P-value, double blind

Signal or image processing

- Cryo-EM resolution enhancement

Data mining

- Complete genome databases

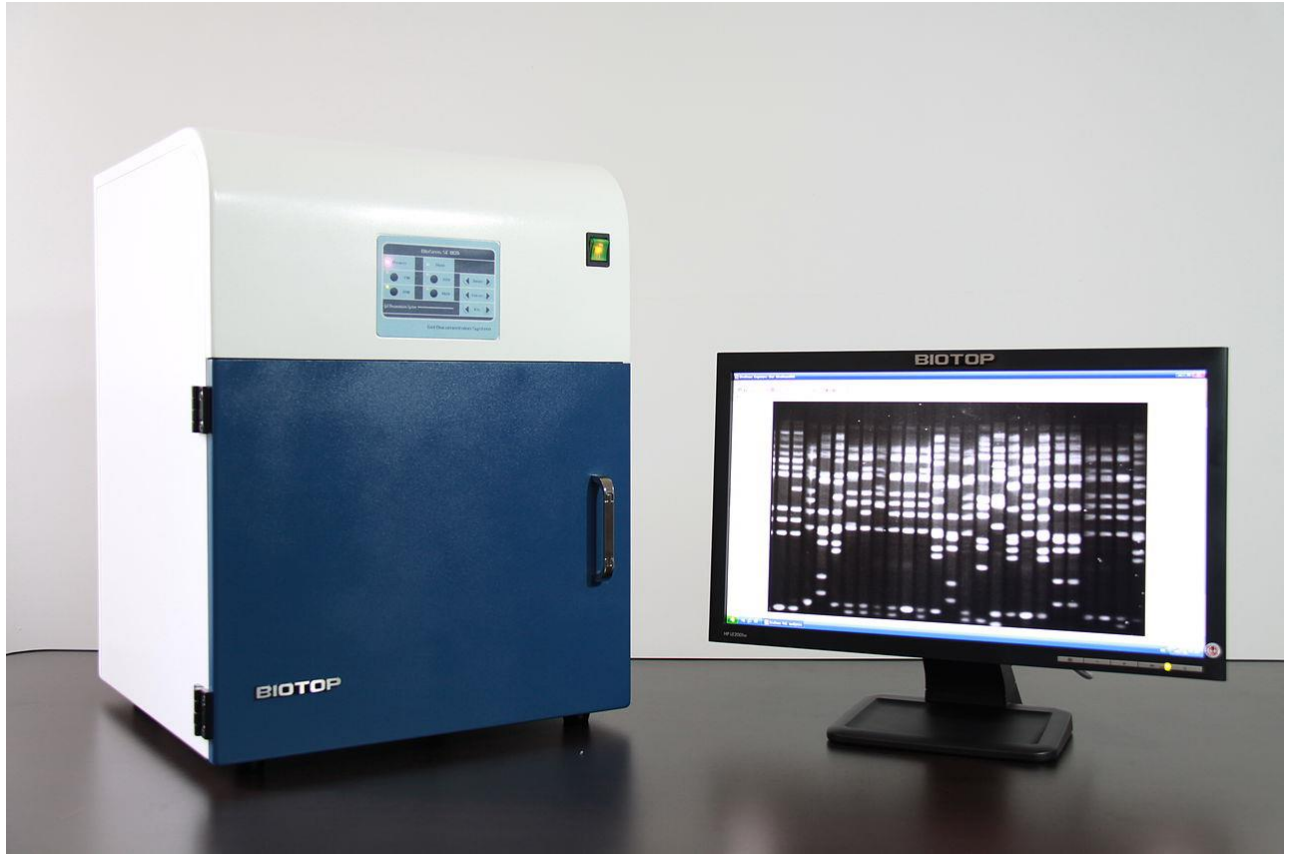
Protein bioinfomatics

Bioinformatic methods related to proteins

no genome data, no patients

Exotic methods

Gel Doc system
image processing



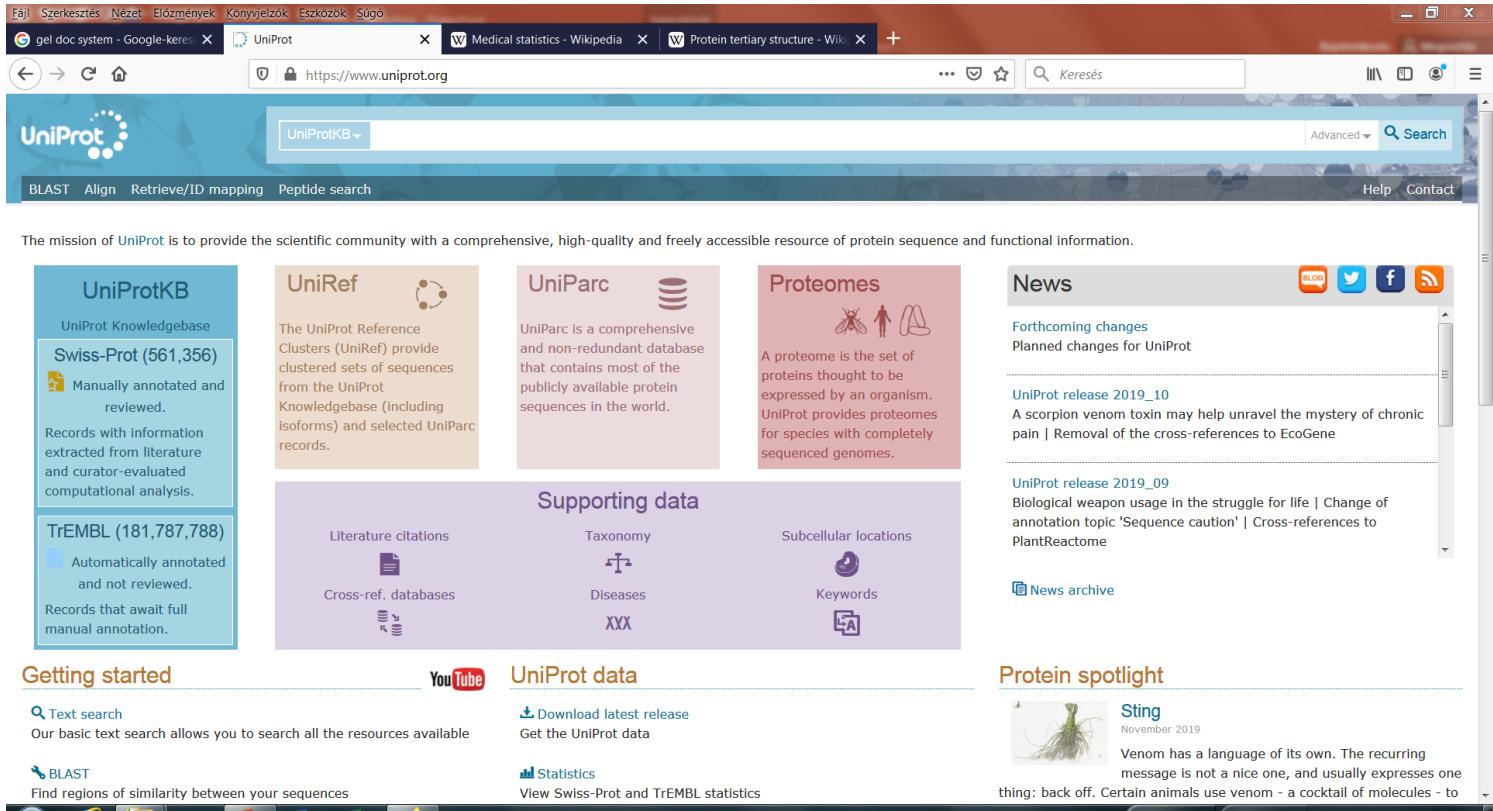
Methods dealing with the amino acid sequences

Primary structure:

Fasta format

Databases

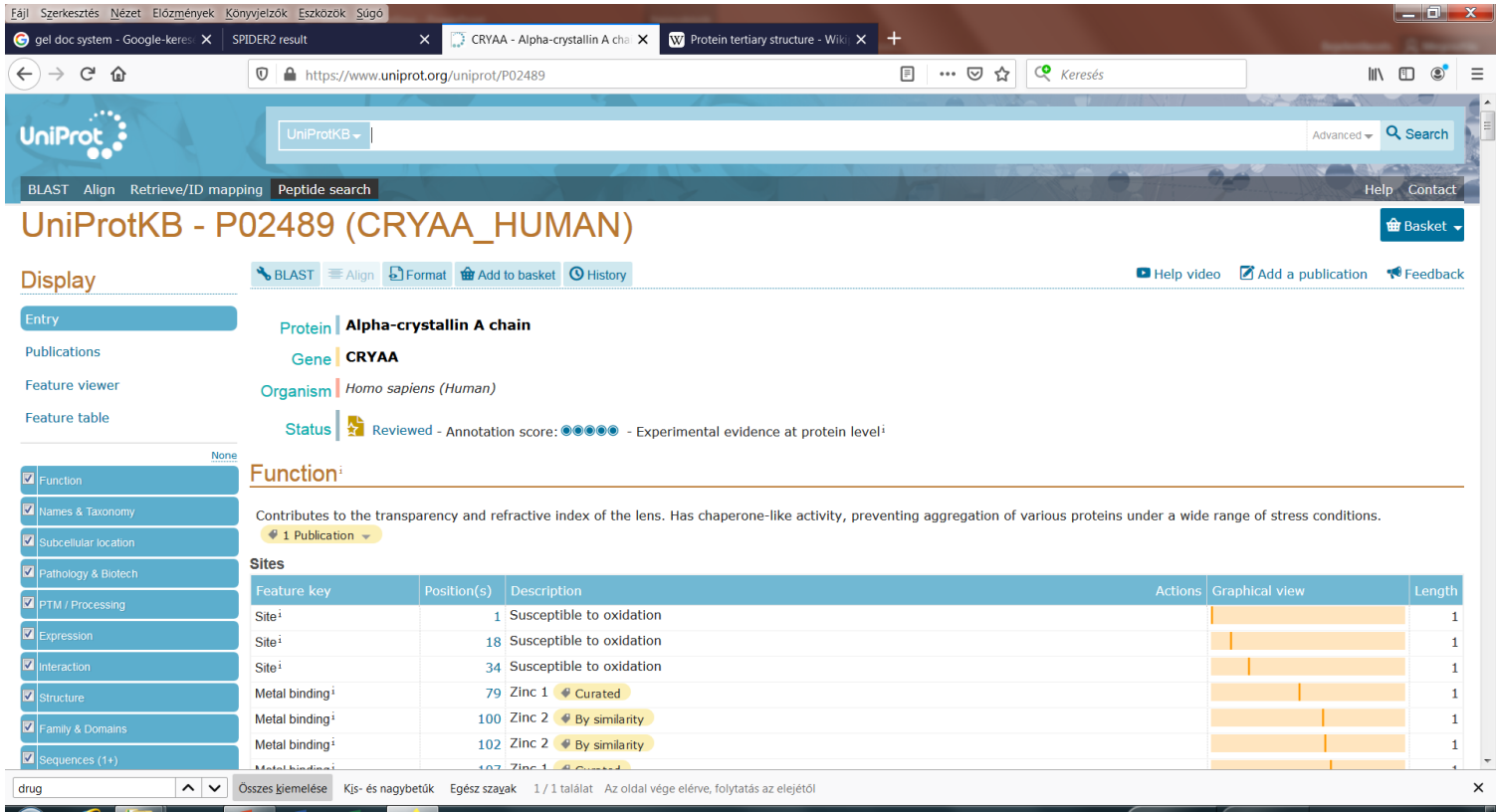
Similiraty searches



Methods dealing with the primary protein structure

Amino acid sequence of alpha-crystallin in Fasta format

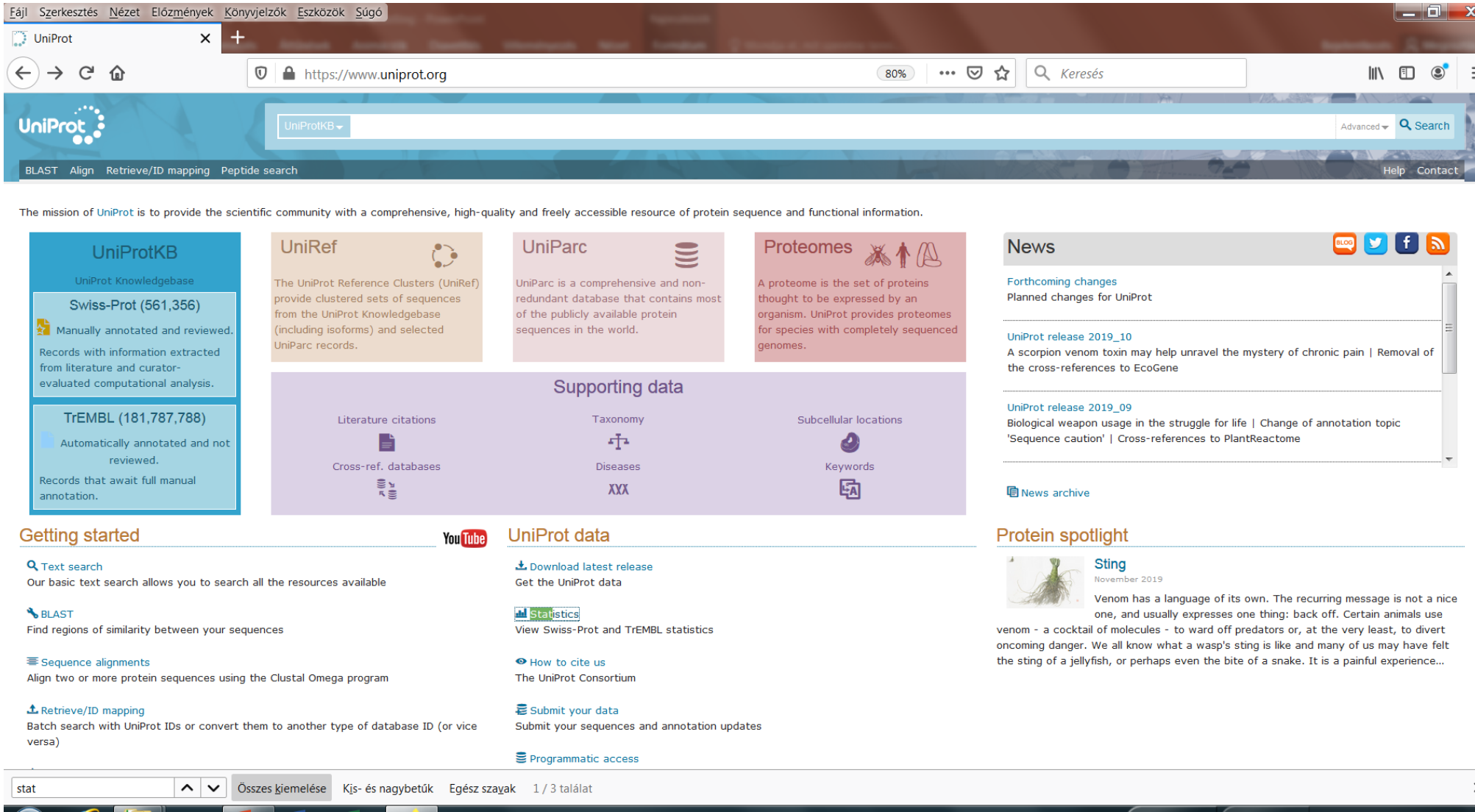
```
>sp|P02489|CRYAA_HUMAN Alpha-crystallin A chain OS=Homo sapiens OX=9606 GN=CRYAA PE=1 SV=2
MDVTIQHPWFKRTLGPFYPSRLFDQFFGEGLFEYDLLPFLSSTISPYYRQSLFRTVLDSG
ISEVRSDRDKFVIFLDVKHFSPEDLTVKVQDDFVEIHGKHNERQDDHGYISREFHRRYRL
PSNVDQSALSCSLSADGMLTFCGPKIQTGLDATHAERAIPVSREEKPTSAPSS
```



Uniprot database

Swissprot: manually annotated

TrEMBL: automatically annotated

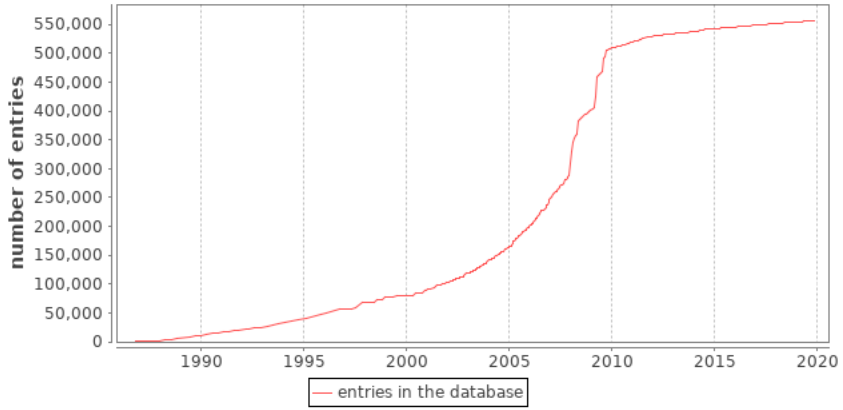


Nov-13, 2019

Introduction

	Number of entries
New entries	181
Updated entries	89,827
Unchanged entries	471,348
Total	561,356
Entries with updated sequences	30
With a fragmented AA sequence	9,203
With known alternative products	25,352

Number of entries in UniProtKB/Swiss-Prot over time



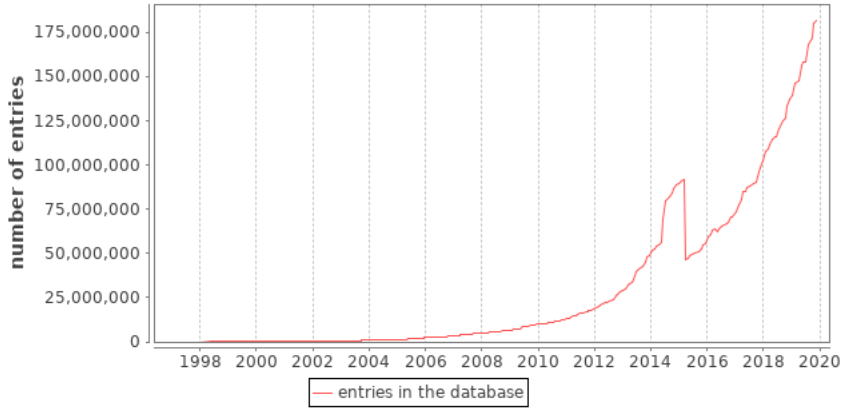
	Protein Existence (PE)	Number of entries
1	Evidence at protein level	101,928
2	Evidence at transcript level	57,282
3	Inferred from homology	386,905
4	Predicted	13,404
5	Uncertain	1,837

Nov-13, 2019

Introduction

	Number of entries
New entries	6,122,937
Updated entries	39,948,743
Unchanged entries	135,716,108
Total	181,787,788
Entries with updated sequences	900
With a fragmented AA sequence	17,497,635
With known alternative products	0

Number of entries in UniProtKB/TrEMBL over time

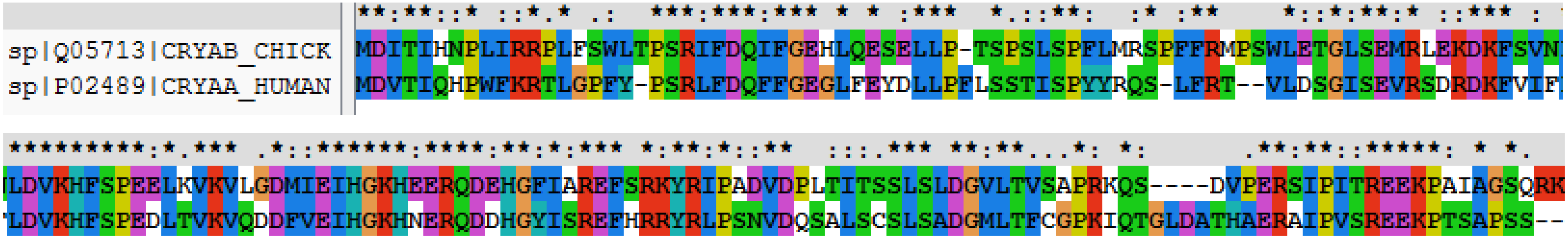


	Protein Existence (PE)	Number of entries
1	Evidence at protein level	151,242
2	Evidence at transcript level	1,297,981
3	Inferred from homology	45,460,655
4	Predicted	134,877,910
5	Uncertain	0

Methods dealing with protein sequences

Similarity of the protein sequences; Clustal program

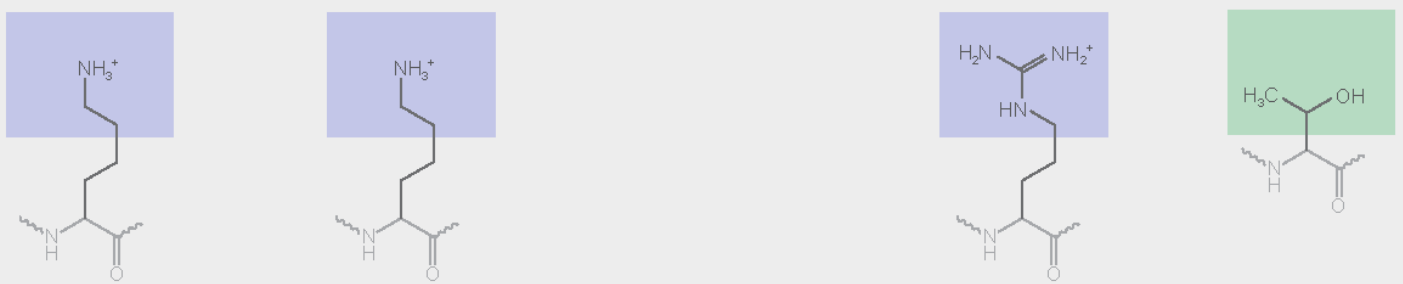
pairwise sequence alignment; neighbor joining



Similiraty matrices

Point accepted Mutations

	No mutation	Point mutations			
		Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	**Lys**	**Lys**	**STOP**	**Arg**	**Thr**



PAM-250 matrix

likelihood of an amino acid of being replaced with and other amino acid through a series of one or more point accepted mutations during evolution, rather two amino acids being aligned due to a coincidence

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	J	Z	X
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0	0	-1	0	-1
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	-1	-3	0	-1
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	2	-3	1	-1
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	3	-3	3	-1
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-4	-5	-5	-1
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	1	-2	3	-1
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	3	-3	3	-1
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	0	-4	0	-1
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	1	-2	2	-1
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-2	3	-2	-1
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-3	5	-3	-1
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	1	-3	0	-1
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-2	3	-2	-1
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-4	2	-5	-1
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	-1	-2	0	-1
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	0	-2	0	-1
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	0	-1	-1	-1
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-5	-3	-6	-1
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-3	-1	-4	-1
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	-2	2	-2	-1
B	0	-1	2	3	-4	1	3	0	1	-2	-3	1	-2	-4	-1	0	0	-5	-3	-2	3	-3	2	-1
J	-1	-3	-3	-3	-5	-2	-3	-4	-2	3	5	-3	3	2	-2	-2	-1	-3	-1	2	-3	5	-2	-1
Z	0	0	1	3	-5	3	3	0	2	-2	-3	0	-2	-5	0	0	-1	-6	-4	-2	2	-2	3	-1
X	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

BLOcks SUbstitution Matrix

for diverse sequences

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

PAM vs BLOSUM

The PAM matrices are based alignments including conserved and variable regions. BLOSUM is based on highly conserved regions, without gaps

BLOSUM procedure uses groups of sequences within not all mutations are counted the same

High numbers in the PAM matrix mean larger evolutionary distance, while larger numbers in the BLOSUM matrix denote higher sequence similarity:
PAM250 is used for more distant sequences than PAM100;
BLOSUM62 is used for closer sequences than BLOSUM50

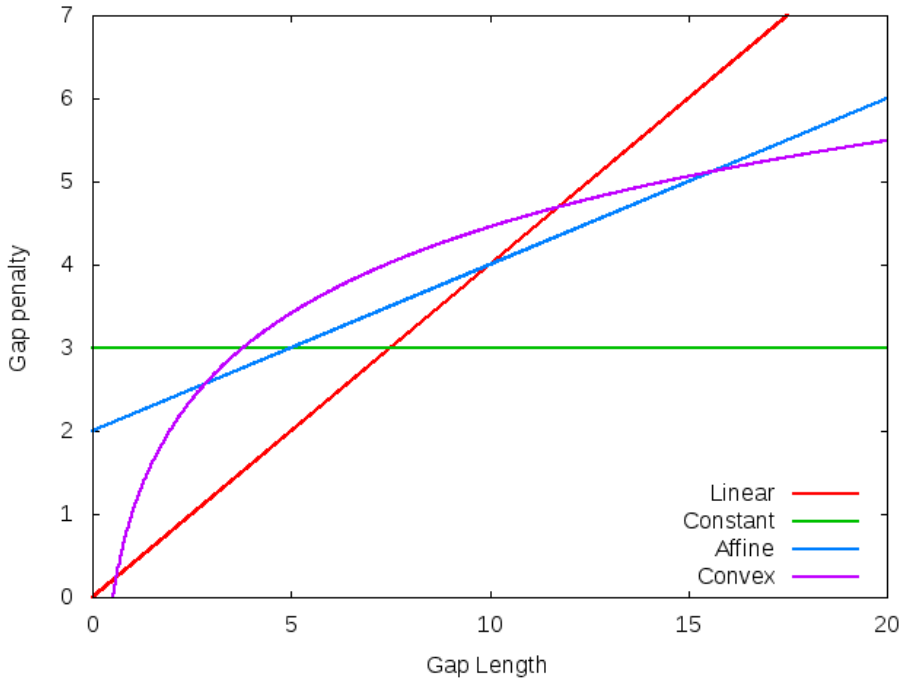
Insertion and deletions

gap opening penalty

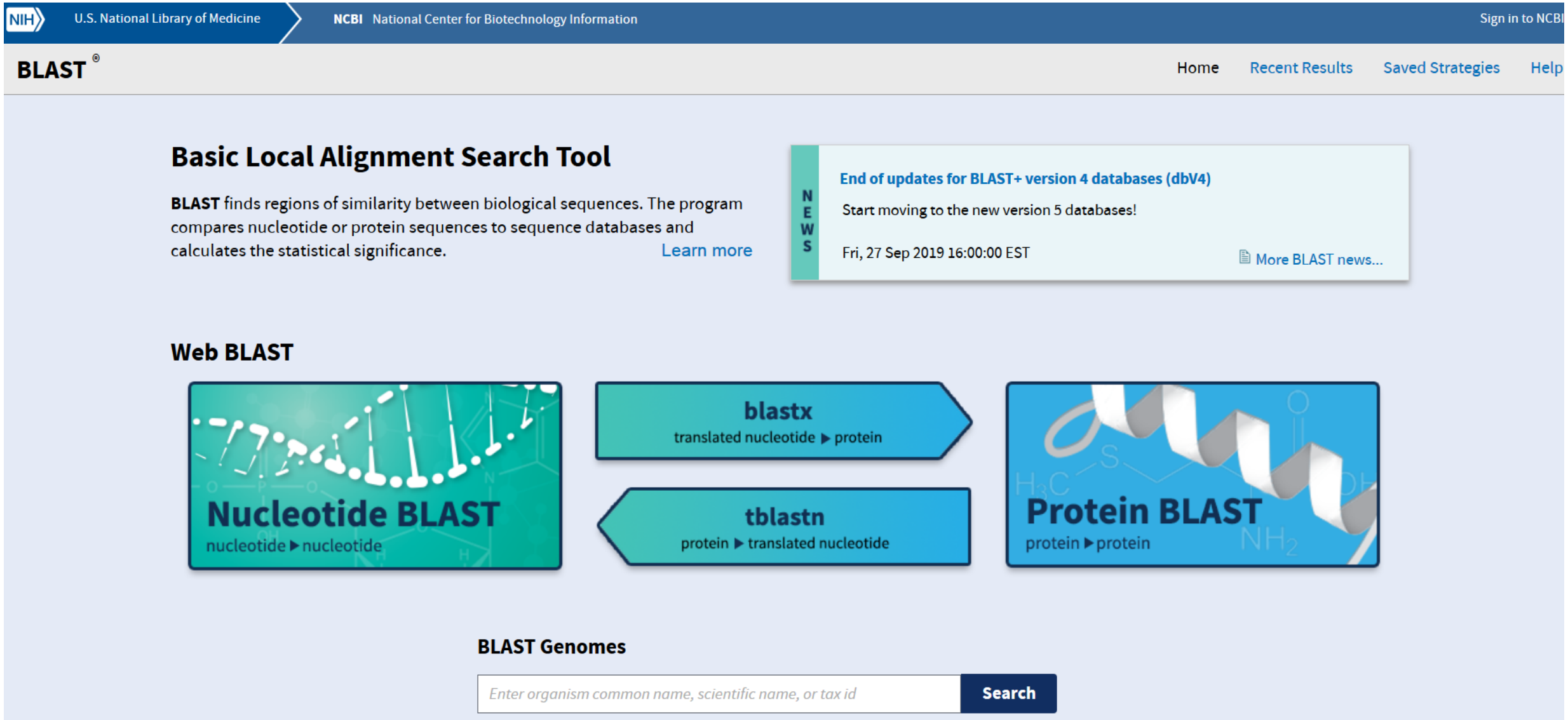
gap extension penalty

constant / linear

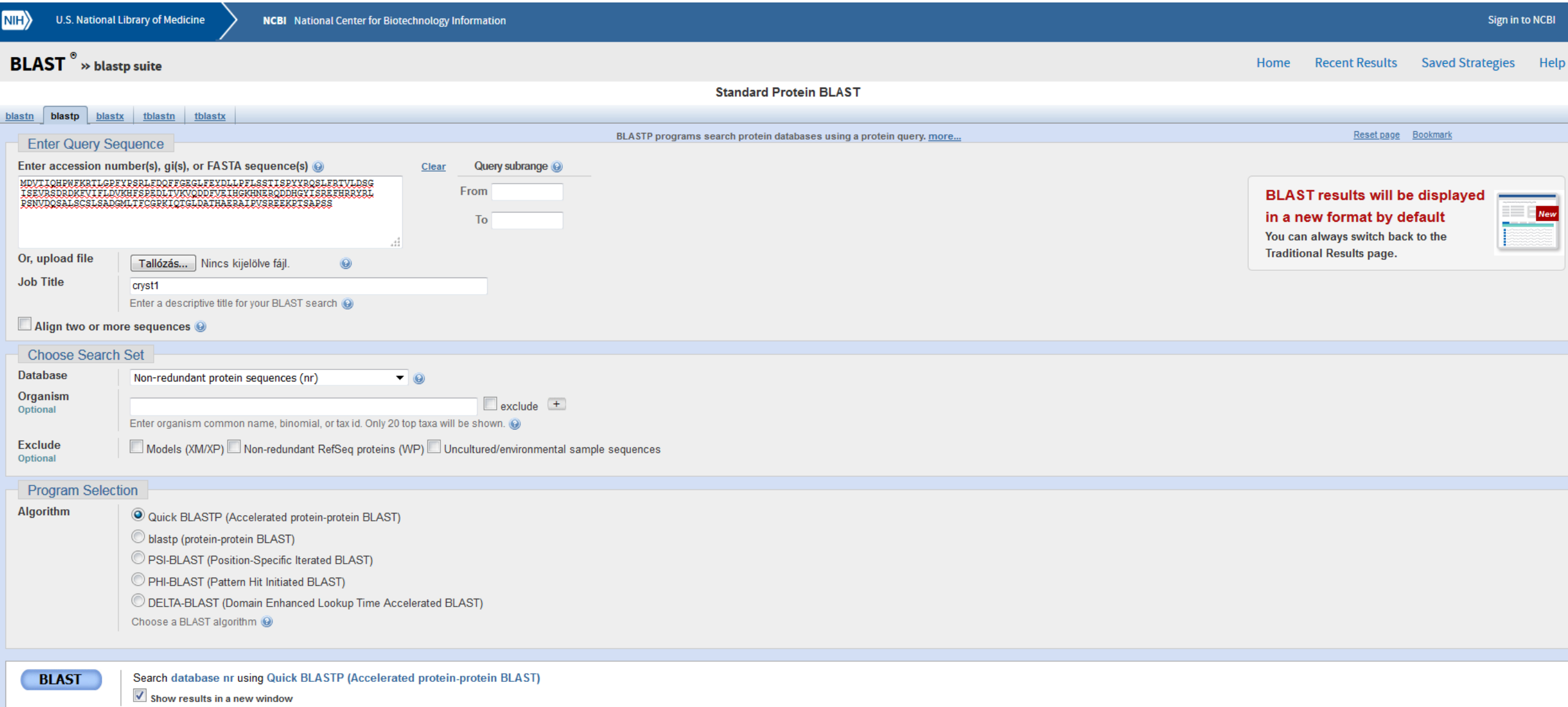
```
CAA37898.1      -----------MSTLEGRGFTE--EQEALVVKSWSAMKPNAGELGLKFFLKIFEIAPSAQ   47
P68871.2        -----------MVHLTPEEKSA-------VTALWG-KV-NVDEVGGEALGRLLVVYPWTQ   40
CAA77743.1      MHSSIVLATVLFVAIASASKTRELCMKSLEHAKVG-TSKEAKQDGIDLYKHMFEHYPAMK   59
AAA29796.1      MHSSIVLATVLFVAIASASKTRELCMKSLEHAKVG-TSKEAKQDGIDLYKHMFEHYPAMK   59
                           :  :     :              .    :. : * .   :::   *  :

CAA37898.1      KLFSFLKDSNVPL--ERNPKLKSHAMSVFLMTCESAVQLRKAGKVTVRESSLKKLGASHF   105
P68871.2        RFFESFGDLSTPDAVMGNPKVKAHGKKVLG-AFS-------DGL----AHLDNLKGTFAT   88
CAA77743.1      KYFKHRENY-TPADVQKDPFFIKQGQNILL-ACHVLCATY-DDR----ETFDAYVGELMA   112
AAA29796.1      KYFKHRENY-TPADVQKDPFFIKQGQNILL-ACHVLCATY-DDR----ETFDAYVGELMA   112
                : *.   :  .*     :* .  :. .::  :          .            *

CAA37898.1      KHGVAD-------EHFEVTKFALLETIKEAVPETWSPEMKNAWGEAYDKLVAAIKLEMKP   158
P68871.2        LSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHK---   145
CAA77743.1      RHE--RDHVKIPNDVWNHFWEHFIEFLG--SKTTLDEPTKHAWQEIGKEFSHEISHHGRH   168
AAA29796.1      RHE--RDHVKVPNDVWNHFWEHFIEFLG--SKTTLDEPTKHAWQEIGKEFSHEISHHGRH   168
                      : :.     ::  :           : *: :    .   :  .
```



Similarity searches: BLAST



Crystallin non-redundant sequence search



Crystallin non-redundant sequence search

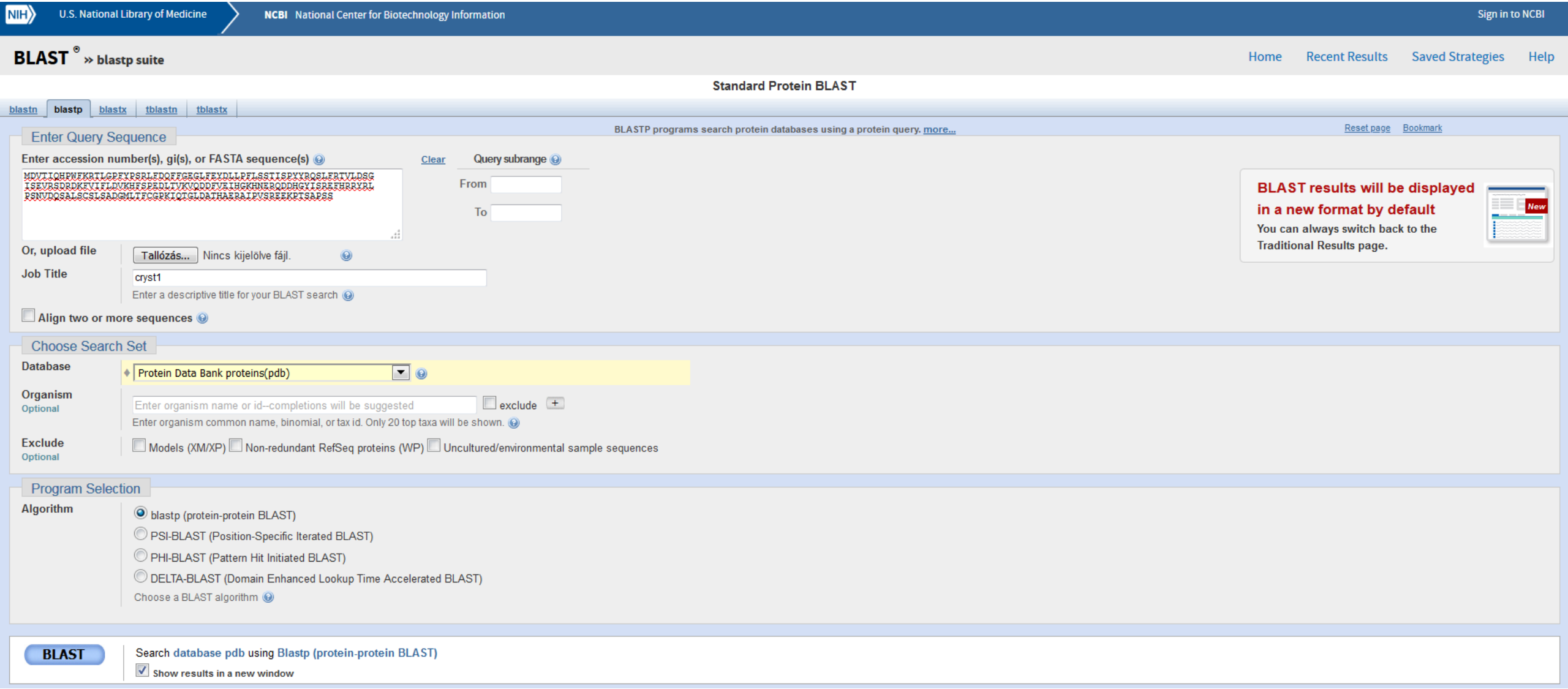
Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Manage Columns Show 100

select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
alpha-crystallin A chain isoform 1 [Homo sapiens]	357	357	100%	2e-124	100.00%	NP_000385.1
alpha-crystallin A chain isoform X1 [Nomascus leucogenys]	353	353	100%	3e-123	98.84%	XP_030661912.1
CRYAA [synthetic construct]	353	353	100%	3e-123	99.42%	AKI71705.1
CRYAA [synthetic construct]	353	353	100%	5e-123	99.42%	AKI71706.1
CRYAA [synthetic construct]	353	353	100%	5e-123	99.42%	AKI71707.1
CRYAA [synthetic construct]	353	353	100%	5e-123	99.42%	AKI71704.1
alpha-crystallin A chain [Aotus nancymaae]	346	346	100%	2e-120	96.53%	XP_012300248.1
RecName: Full=Alpha-crystallin A chain [Loxodonta africana]	346	346	100%	3e-120	95.38%	P02498.1
PREDICTED: alpha-crystallin A chain [Saimiri boliviensis boliviensis]	345	345	100%	7e-120	95.95%	XP_003927618.1
PREDICTED: alpha-crystallin A chain [Macaca fascicularis]	345	345	100%	8e-120	97.69%	XP_005548643.1
alpha-crystallin A chain [Oryctolagus cuniculus]	345	345	100%	8e-120	95.95%	NP_001075875.2
PREDICTED: alpha-crystallin A chain [Chlorocebus sabaeus]	344	344	100%	1e-119	97.69%	XP_007967599.1
Heat shock protein beta-4 [Macaca mulatta]	345	345	100%	1e-119	97.69%	EHH16937.1
alpha-crystallin A chain [Carlito syrichta]	343	343	100%	2e-119	95.95%	XP_008069166.1
alpha-crystallin A chain [Piliocolobus tephrosceles]	346	346	100%	3e-119	97.69%	XP_026303245.1
alpha-crystallin A chain isoform X3 [Octodon degus]	343	343	100%	4e-119	95.38%	XP_004631834.1
alpha-A-crystallin [Oryctolagus cuniculus]	343	343	100%	5e-119	95.38%	CAA64668.1
crystallin alphaA [Macaca mulatta]	342	342	100%	6e-119	97.11%	751000B
LOW QUALITY PROTEIN: alpha-crystallin A chain-like [Rhinopithecus roxellana]	342	342	100%	6e-119	97.11%	XP_030782449.1
PREDICTED: alpha-crystallin A chain isoform X3 [Peromyscus maniculatus bairdii]	342	342	100%	8e-119	94.80%	XP_006991023.1
alpha-crystallin A chain [Loxodonta africana]	344	344	100%	8e-119	94.80%	XP_003419079.2
Heat shock protein beta-4 [Macaca fascicularis]	343	343	100%	9e-119	97.11%	EHH51832.1
PREDICTED: alpha-crystallin A chain [Cebus capucinus imitator]	341	341	100%	2e-118	94.80%	XP_017394045.1
alpha-crystallin A chain isoform X2 [Microcebus murinus]	341	341	100%	2e-118	94.80%	XP_012634559.1

Crystallin PDB search



Crystallin PDB search

Descriptions | Graphic Summary | Alignments | Taxonomy

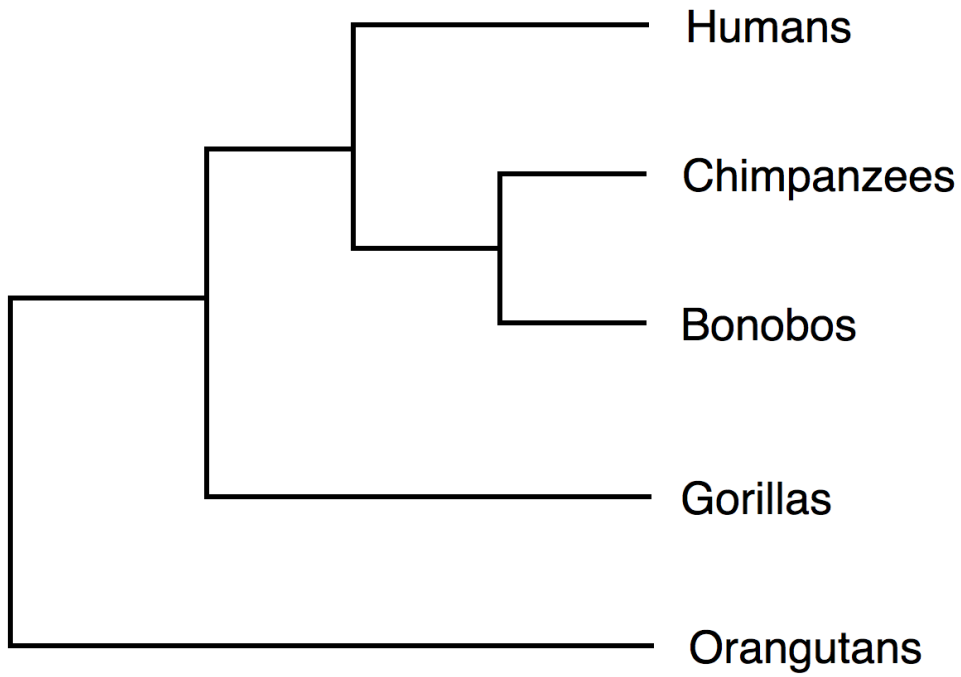
Sequences producing significant alignments Download Manage Columns Show 100

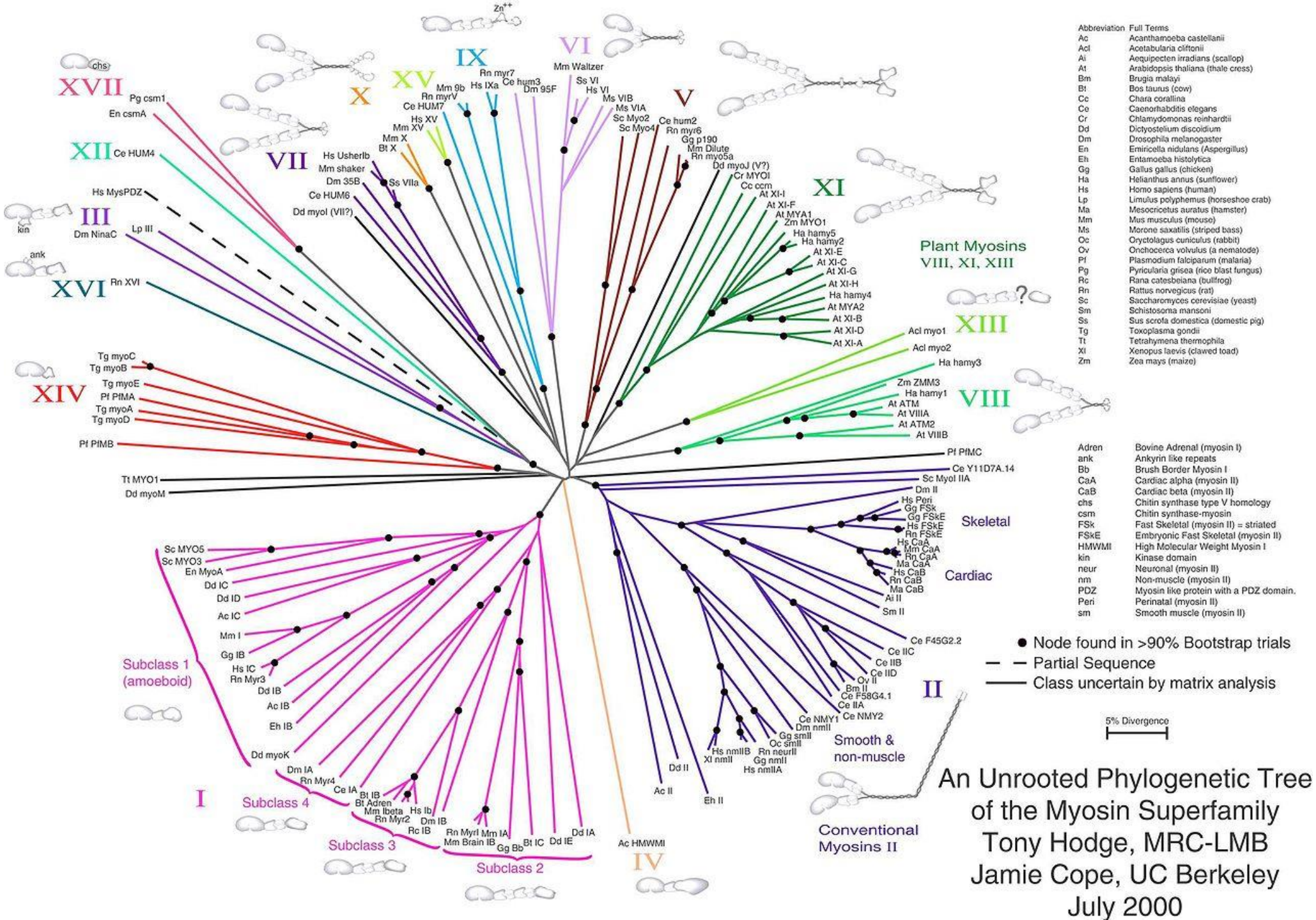
select all *27 sequences selected* GenPept Graphics Distance tree of results Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Chain A, Alpha-crystallin A chain [Bos taurus]	203	203	60%	3e-68	93.33%	3L1E_A
Chain A, Alpha-crystallin A chain [Bos taurus]	197	197	58%	5e-66	93.14%	3L1F_A
Chain A, Alpha-crystallin B chain [Homo sapiens]	189	189	98%	1e-61	54.49%	2KLR_A
Chain A, Alpha A crystallin [Danio rerio]	175	175	61%	3e-57	73.83%	3N3E_A
Chain C, Heat Shock Protein Beta-6 [Homo sapiens]	139	139	81%	1e-42	44.14%	5LTW_C
Chain A, Alpha-crystallin B chain [Homo sapiens]	120	120	51%	8e-36	59.55%	2N0K_A
Chain A, Human Alphab Crystallin [Homo sapiens]	119	119	57%	5e-35	54.55%	3L1G_A
Chain A, Alpha-crystallin B chain [Homo sapiens]	118	118	51%	5e-35	58.43%	6BP9_A
Chain A, ALPHA-CRYSTALLIN B CHAIN [Homo sapiens]	118	118	51%	7e-35	58.43%	2WJ7_A
Chain A, ALPHA-CRYSTALLIN B [Homo sapiens]	115	115	51%	1e-33	57.30%	2Y22_A
Chain A, Alpha-crystallin B Chain [Homo sapiens]	115	115	51%	2e-33	56.18%	2Y1Z_A
Chain A, Alpha-crystallin B Chain [Homo sapiens]	114	114	49%	2e-33	58.82%	4M5S_A
Chain A, Heat shock protein beta-1 [Homo sapiens]	115	115	86%	2e-32	41.18%	6DV5_A
Chain A, ALPHA-CRYSTALLIN B CHAIN [Homo sapiens]	111	111	47%	4e-32	58.54%	2Y1Y_A
Chain A, Alpha-crystallin B chain [Homo sapiens]	110	110	49%	6e-32	57.65%	4M5T_A
Chain A, Heat shock protein beta-1 [Homo sapiens]	105	105	50%	2e-29	55.68%	2N3J_A
Chain A, HEAT SHOCK PROTEIN BETA-6 [Rattus norvegicus]	103	103	47%	8e-29	52.44%	2WJ5_A
Chain A, Heat Shock Protein Beta-1 [Homo sapiens]	101	101	49%	3e-28	55.29%	4MJH_A
Chain A, Heat shock protein beta-1 [Homo sapiens]	101	101	49%	4e-28	55.29%	6GJH_A
Chain A, Heat Shock Protein Beta-6 [Homo sapiens]	100	100	46%	9e-28	50.62%	4JUS_A
Chain A, Heat Shock Protein Beta-6 [Homo sapiens]	99.0	99.0	46%	6e-27	49.38%	4JUT_A
Chain A, Heat Shock Protein Beta-6 [Homo sapiens]	95.5	95.5	41%	5e-26	54.93%	5LUM_A
Chain D, Heat shock protein beta-2 [Homo sapiens]	97.4	97.4	76%	2e-25	40.30%	6F2R_D

Phylogenetic trees

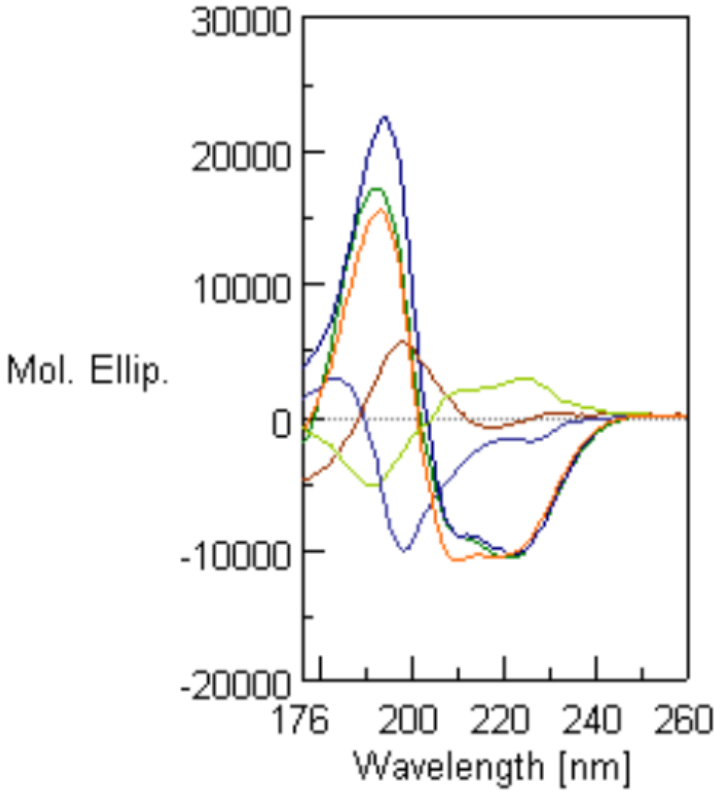
evolutionary relationships between species

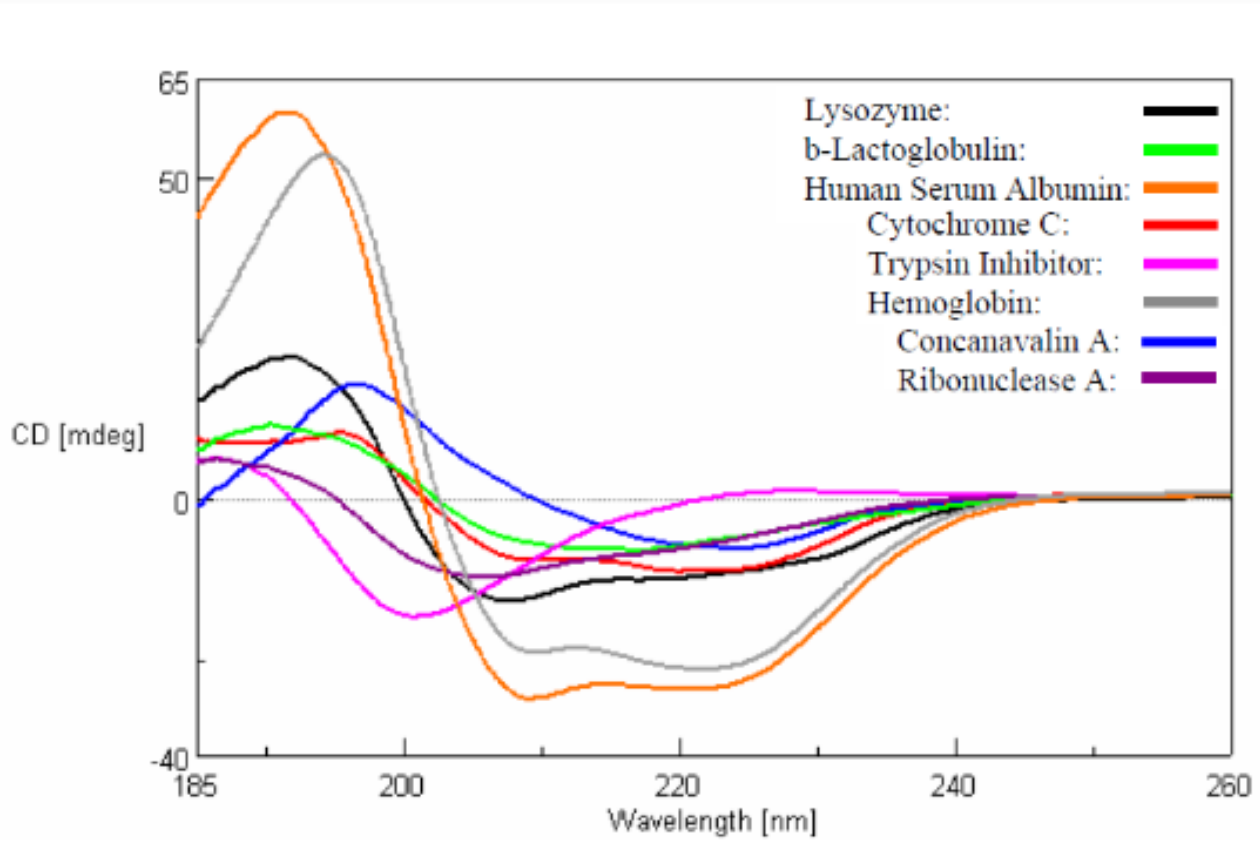



Zn++
chs
XVII
Pg csm1
En csmA
XII Ce HUM4
kin
III
Dm NinaC
Lp III
Hs MysPDZ
ank
XVI
Rn XVI
XIV
Tg myoC
Tg myoB
Tg myoE
Pf PfMA
Tg myoA
Tg myoD
Pf PfMB
Tt MYO1
Dd myoM
Sc MYO5
Sc MYO3
En MyoA
Dd IC
Dd ID
Ac IC
Mm I
Gg IB
Hs IC
Rn Myr3
Dd IB
Ac IB
Eh IB
Dd myoK
Subclass 1
(amoeboid)
I
Dm IA
Rn Myr4
Ce IA
Bt IB
Bt Adren
Mm Ibeta
Hs Ib
Rn Myr2
Dm IB
Rc IB
Subclass 4
Subclass 3
Rn MyrI
Mm Brain IB
Mm IA
Gg Bb
Bt IC
Dd IE
Dd IA
Subclass 2
Ac HMWMI
IV
Ac II
Dd II
Eh II
II
Ce F45G2.2
Ce IIC
Ce IIB
Ce IID
Ov II
Bm II
Ce F58G4.1
Ce IIA
Ce NMY1
Ce NMY2
Dm nmII
Gg smII
Oc smII
Rn neurII
Gg nmII
Hs nmIIB
Xl nmII
Hs nmIIA
Smooth &
non-muscle
Conventional
Myosins II
Ce Y11D7A.14
Pf PfMC
Sc MyoI IIA
Dm II
Hs Peri
Gg FSk
Gg FSkE
Hs FSkE
Rn FSkE
Hs CaA
Mm CaA
Rn CaA
Ma CaA
Hs CaB
Rn CaB
Ma CaB
Ai II
Sm II
Skeletal
Cardiac
IX
Rn myr7
Hs IXa
Ce hum3
Dm 95F
VI
Mm Waltzer
Ss VI
Hs VI
Ms VIB
Ms VIA
Sc Myo2
Sc Myo4
V
Ce hum2
Rn myr6
Gg p190
Mm Dilute
Rn myo5a
Dd myoJ (V?)
Cr MYOI
Cc ccm
At XI-I
At XI-F
At MYA1
Zm MYO1
Ha hamy5
Ha hamy2
At XI-E
At XI-C
At XI-G
At XI-H
Ha hamy4
At MYA2
At XI-B
At XI-D
At XI-A
XI
Plant Myosins
VIII, XI, XIII
?
XIII
Acl myo1
Acl myo2
Ha hamy3
Zm ZMM3
Ha hamy1
At ATM
At VIIIA
At ATM2
At VIIIB
VIII
X
XV
Mm 9b
Rn myrV
Ce HUM7
Hs XV
Mm XV
Mm X
Bt X
VII
Hs Usherlb
Mm shaker
Dm 35B
Ss VIIa
Ce HUM6
Dd myoI (VII?)
Abbreviation Full Terms
Ac Acanthamoeba castellanii
Acl Acetabularia cliftonii
Ai Aequipecten irradians (scallop)
At Arabidopsis thaliana (thale cress)
Bm Brugia malayi
Bt Bos taurus (cow)
Cc Chara corallina
Ce Caenorhabditis elegans
Cr Chlamydomonas reinhardtii
Dd Dictyostelium discoidium
Dm Drosophila melanogaster
En Emericella nidulans (Aspergillus)
Eh Entamoeba histolytica
Gg Gallus gallus (chicken)
Ha Helianthus annus (sunflower)
Hs Homo sapiens (human)
Lp Limulus polyphemus (horseshoe crab)
Ma Mesocricetus auratus (hamster)
Mm Mus musculus (mouse)
Ms Morone saxatilis (striped bass)
Oc Oryctolagus cuniculus (rabbit)
Ov Onchocerca volvulus (a nematode)
Pf Plasmodium falciparum (malaria)
Pg Pyricularia grisea (rice blast fungus)
Rc Rana catesbeiana (bullfrog)
Rn Rattus norvegicus (rat)
Sc Saccharomyces cerevisiae (yeast)
Sm Schistosoma mansoni
Ss Sus scrofa domestica (domestic pig)
Tg Toxoplasma gondii
Tt Tetrahymena thermophila
Xl Xenopus laevis (clawed toad)
Zm Zea mays (maize)
Adren Bovine Adrenal (myosin I)
ank Ankyrin like repeats
Bb Brush Border Myosin I
CaA Cardiac alpha (myosin II)
CaB Cardiac beta (myosin II)
chs Chitin synthase type V homology
csm Chitin synthase-myosin
FSk Fast Skeletal (myosin II) = striated
FSkE Embryonic Fast Skeletal (myosin II)
HMWMI High Molecular Weight Myosin I
kin Kinase domain
neur Neuronal (myosin II)
nm Non-muscle (myosin II)
PDZ Myosin like protein with a PDZ domain.
Peri Perinatal (myosin II)
sm Smooth muscle (myosin II)
● Node found in >90% Bootstrap trials
– – Partial Sequence
—— Class uncertain by matrix analysis
5% Divergence
An Unrooted Phylogenetic Tree
of the Myosin Superfamily
Tony Hodge, MRC-LMB
Jamie Cope, UC Berkeley
July 2000

Secondary structure of proteins

Decomposition of CD and FTIR spectra





		α-Helix %	β-Sheet %	Turn %	Other %
Lysozyme	PLS	42.8	0.4	24.4	32.4
	X-ray[1]	41	4	19	35
Cytochrome C	PLS	42.6	3.1	18.1	36.2
	X-ray[1]	42	8	9	42
Concanavalin A	PLS	5.1	44.6	13.9	36.4
	X-ray[1]	2	36	12	49
β-Lactoglobulin	PLS	17.8	35.5	12.3	34.4
	X-ray[1]	13	34	13	41
Trypsin inhibitor	PLS	13.9	25.3	17.3	43.5
	X-ray[2]	2	33	10	55
Ribonuclease A	PLS	21.5	14.7	22.4	41.4
	X-ray[1]	22	19	11	48
Human Serum Albumin (HSA)	PLS	66.8	1.3	8.2	23.7
	X-ray[2]	72	0	8	29
Hemoglobin	PLS	61.1	0	18	20.9
	X-ray[1]	75	0	10	15

DSSP

```
==== Secondary Structure Definition by the program DSSP, CMBI version by M.L. Hekkelman/2010-10-21 ==== DATE=2019-11-22        .
REFERENCE W. KABSCH AND C.SANDER, BIOPOLYMERS 22 (1983) 2577-2637                                                            .
HEADER    CHAPERONE                               08-JUL-09   2KLR                                                        .
COMPND    2 MOLECULE: ALPHA-CRYSTALLIN B CHAIN;                                                                          .
SOURCE    2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;                                                                           .
AUTHOR    S.JEHLE,P.RAJAGOPAL,S.MARKOVIC,B.BARDIAUX,R.KUEHNE,V.A.HIGMAN,                                                 .
  164  2  0  0  0 TOTAL NUMBER OF RESIDUES, NUMBER OF CHAINS, NUMBER OF SS-BRIDGES(TOTAL,INTRACHAIN,INTERCHAIN)          .
  9783.4   ACCESSIBLE SURFACE OF PROTEIN (ANGSTROM**2)                                                                  .
  110 67.1   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(J)  , SAME NUMBER PER 100 RESIDUES                        .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS IN     PARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES                       .
   78 47.6   TOTAL NUMBER OF HYDROGEN BONDS IN ANTIPARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES                       .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-5), SAME NUMBER PER 100 RESIDUES                       .
    2  1.2   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-4), SAME NUMBER PER 100 RESIDUES                       .
    4  2.4   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-3), SAME NUMBER PER 100 RESIDUES                       .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-2), SAME NUMBER PER 100 RESIDUES                       .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-1), SAME NUMBER PER 100 RESIDUES                       .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+0), SAME NUMBER PER 100 RESIDUES                       .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+1), SAME NUMBER PER 100 RESIDUES                       .
   18 11.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+2), SAME NUMBER PER 100 RESIDUES                       .
   12  7.3   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+3), SAME NUMBER PER 100 RESIDUES                       .
    4  2.4   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+4), SAME NUMBER PER 100 RESIDUES                       .
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+5), SAME NUMBER PER 100 RESIDUES                       .
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30     *** HISTOGRAMS OF ***     .
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    RESIDUES PER ALPHA HELIX   .
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    PARALLEL BRIDGES PER LADDER .
 0  0  0  2  0  4  0  0  0  0  1  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    ANTIPARALLEL BRIDGES PER LADDER .
 0  2  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    LADDERS PER SHEET          .
 #  RESIDUE AA STRUCTURE BP1 BP2  ACC     N-H-->O    O-->H-N    N-H-->O    O-->H-N    TCO  KAPPA ALPHA  PHI   PSI    X-CA   Y-CA   Z-CA
   1   69 A R              0   0  223      0, 0.0     0, 0.0     0, 0.0     0, 0.0   0.000 360.0 360.0 360.0 -37.4  -37.9    7.6   -0.7
   2   70 A L        -     0   0   70      2,-0.1     0, 0.0    78,-0.0     0, 0.0   0.973 360.0 -59.8  56.7  51.2  -34.2    7.5   -1.5
   3   71 A E        -     0   0   78      1,-0.2     2,-1.0    77,-0.1    78,-0.4   0.912  52.6-130.1  42.4 110.0  -34.8    4.2   -3.1
   4   72 A K        +     0   0  179     76,-0.1     2,-0.3     2,-0.1    -1,-0.2  -0.708  67.1 103.1 -86.9 106.3  -36.2    1.4   -1.0
   5   73 A D  S     S-    0   0   81     -2,-1.0     2,-0.8     2,-0.1    76,-0.7  -0.802  70.8-118.7-177.6 138.1  -33.9   -1.5   -1.7
   6   74 A R  E     -A   80   0A 214     -2,-0.3     2,-0.4    74,-0.2    -2,-0.1  -0.838  37.0-170.7 -94.4 117.6  -31.2   -2.9    0.3
   7   75 A F  E     -A   79   0A  45     72,-2.6    72,-1.8    -2,-0.8     2,-0.4  -0.821   6.1-170.1-110.7 142.6  -28.0   -2.6   -1.6
   8   76 A S  E     -A   78   0A  65     -2,-0.4     2,-0.4    70,-0.2    70,-0.2  -0.994  10.2-178.9-137.6 131.6  -24.7   -4.1   -0.8
   9   77 A V  E     -A   77   0A  15     68,-2.8    68,-3.2    -2,-0.4     2,-0.4  -0.961  21.7-143.1-118.9 141.7  -21.3   -3.6   -2.2
  10   78 A N  E     -A   76   0A  64     -2,-0.4     2,-0.5    66,-0.2    66,-0.2  -0.873   8.8-161.8-111.8 143.8  -18.5   -5.6   -0.8
  11   79 A L  E     -A   75   0A   0     64,-3.7    64,-3.1    -2,-0.4     2,-0.5  -0.989  16.1-153.7-120.0 127.6  -15.0   -4.4   -0.3
  12   80 A D        +     0   0   30     -2,-0.5     2,-0.5    62,-0.2    62,-0.2  -0.916  23.0 175.9-117.2 127.3  -12.5   -7.1    0.1
  13   81 A V        -     0   0    2     -2,-0.5     3,-0.3    60,-0.3    -2,-0.1  -0.997  27.8-144.1-118.9 116.6   -9.2   -7.0    1.9
  14   82 A K  S     S-    0   0  105     -2,-0.5     2,-0.3     1,-0.3    -1,-0.1   0.876  77.2  -8.8 -54.4 -54.3   -7.9  -10.5    1.7
  15   83 A H  S     S+    0   0  113     58,-0.1     2,-0.3    -3,-0.1    -1,-0.3  -0.852  85.3 117.4-150.4 122.0   -6.3  -10.6    5.1
  16   84 A F        -     0   0   12     56,-1.4    58,-0.1    -3,-0.3    -3,-0.1  -0.902  55.5-106.8-160.4 171.6   -5.8   -7.7    7.5
  17   85 A S    >   -     0   0   60     -2,-0.3     3,-2.6    56,-0.1    55,-0.1  -0.921  30.1-116.1-112.9 141.1   -7.0   -6.9   10.9
  18   86 A P  G >  S+     0   0   40      0, 0.0     3,-1.3     0, 0.0    -1,-0.1   0.676 111.8  61.3 -48.9 -28.3   -9.7   -4.2   11.6
  19   87 A E  G 3  S+     0   0  170      1,-0.3     0, 0.0     3,-0.0     0, 0.0   0.745 102.7  51.8 -74.0 -19.8   -7.4   -1.9   13.5
  20   88 A E  G <  S+     0   0   70     -3,-2.6    15,-2.3    14,-0.0     2,-0.3   0.154 101.6  90.0 -98.3  18.1   -5.2   -1.6   10.4
  21   89 A L  E <   -C   34   0B  40     -3,-1.3     2,-0.4    13,-0.2    13,-0.2  -0.861  49.3-175.7-120.2 148.0   -8.3   -0.6    8.4
  22   90 A K  E     -C   33   0B 129     11,-2.4    11,-3.4    -2,-0.3     2,-0.5  -0.974  10.2-160.5-138.2 124.7  -10.0    2.6    7.6
  23   91 A V  E     -C   32   0B  54     -2,-0.4     2,-0.4     9,-0.2     9,-0.2  -0.936  12.4-174.2-109.8 129.5  -13.2    2.6    5.7
  24   92 A K  E     -C   31   0B  73      7,-3.0     7,-3.1    -2,-0.5     2,-0.5  -0.984  15.1-176.4-130.0 130.7  -14.1    5.8    4.0
```

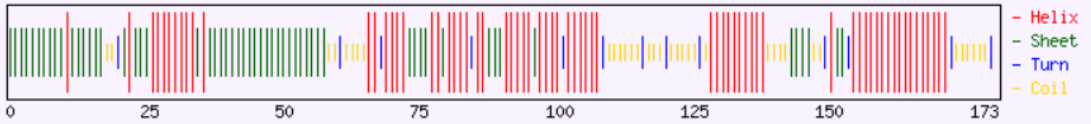
Secondary structure prediction

Chou-Fasman method ~55% accuracy

Amino Acid	(α-Helix)	P (β-Strand)	P (Turn)
Alanine	1.42	0.83	0.66
Arginine	0.98	0.93	0.95
Asparagine	0.67	0.89	1.56
Aspartic acid	1.01	0.54	1.46
Cysteine	0.70	1.19	1.19
Glutamic acid	1.51	0.37	0.74
Glutamine	1.11	1.11	0.98
Glycine	0.57	0.75	1.56
Histidine	1.00	0.87	0.95
Isoleucine	1.08	1.60	0.47
Leucine	1.21	1.30	0.59
Lysine	1.14	0.74	1.01
Methionine	1.45	1.05	0.60
Phenylalanine	1.13	1.38	0.60
Proline	0.57	0.55	1.52
Serine	0.77	0.75	1.43
Threonine	0.83	1.19	0.96
Tryptophan	0.83	1.19	0.96
Tyrosine	0.69	1.47	1.14
Valine	1.06	1.70	0.50

Target Sequence:

```
         10         20         30         40         50         60         70
MDVTIQHPWF KRTLGPFYPS RLFDQFFGEG LFEYDLLPFL SSTISPYYRQ SLFRTVLDSG ISEVRSDRDK

         80         90        100        110        120        130        140
FVIFLDVKHF SPEDLTVKVQ DDFVEIHGKH NERQDDHGYI SREFHRRYRL PSNVDQSALS CSLSADGMLT

        150        160        170
FCGPKIQTGL DATHAERAIP VSREEKPTSA PSS
```



Secondary Structure:

```
                 *         *         *         *         *         *
Query 1   MDVTIQHPWFKRTLGPFYPSRLFDQFFGEGLFEYDLLPFLSSTISPYYRQSLFRTVLDSGISEVRSDRDK 70
Helix 1   HHHHHHHHHHHHH         HHHHHHHHHHHHHHHHHHHHH         HH           HHHHHHH 70
Sheet 1   EEEEEEEEEEEEEEEEE   EEEEEE    EEEEEEEEEEEEEEEEEEEEEEEEE                   70
Turns 1         T    T       T        T           T        T       T      T  T     70
Struc 1   EEEEEEEEEEHEEEEEECCTEHEEEHHHHHHHHHEHEEEEEEEEEEEEEEEEEEEEEEECCTCCCCHHTHHHH 70

                 *         *         *         *         *         *
Query 71  FVIFLDVKHFSPEDLTVKVQDDFVEIHGKHNERQDDHGYISREFHRRYRLPSNVDQSALSCSLSADGMLT 140
Helix 71  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH                    HHHHHHHHHH      140
Sheet 71  EEEEEEE       EEEEEEEEEE                                          EEE 140
Turns 71              TT       T       T   T  T      T   T      T   T              140
Struc 71  EEEEHHEHHHHTHHEEEHHHHHEHHHHTHHHHHHTCCCCCCTCCCTCCCCCTCHHHHHHHHHHCCCCEEE 140

                 *         *         *
Query 141 FCGPKIQTGLDATHAERAIPVSREEKPTSAPSS 173
Helix 141     HHHHHHHHHHHHHHHHHHHHH         173
Sheet 141 E    EEEE                         173
Turns 141    T   T         T     T  T      T 173
Struc 141 ECCTHEETHHHHHHHHHHHHHHHHHTCCCCCCT 173


Total Residues: H: 108   E: 74    T: 25
       Percent: H: 62.4  E: 42.8  T: 14.5
```

Other secondary structure predictors

	Accuracy
Chou-Fasman: single sequence, single residue	~55%
GOR: single sequence, multiple residues	~65%
PHD: neural network, multiple sequences	~71%

Redox state Prediction

Intracellular reducing

~82% accur.

manaslu.fiserlab.org/cysredox.html

Keresés

CYSREDOX

Predicting the redox state of cysteins in proteins from multiple sequence alignments

A. Fiser & I. Simon, Bioinformatics, *16*, 291-325 (2000)

A. Fiser & I. Simon, Methods Enzym., *353*, 10-21 (2002)

Paste your protein sequence in the window ... (use one letter codes)

or upload your sequence file: Tallózás... Nincs kijelölve fájl.

Cutoff for Cystein covalent state prediction 1.27 (range (0.00;unlimited))

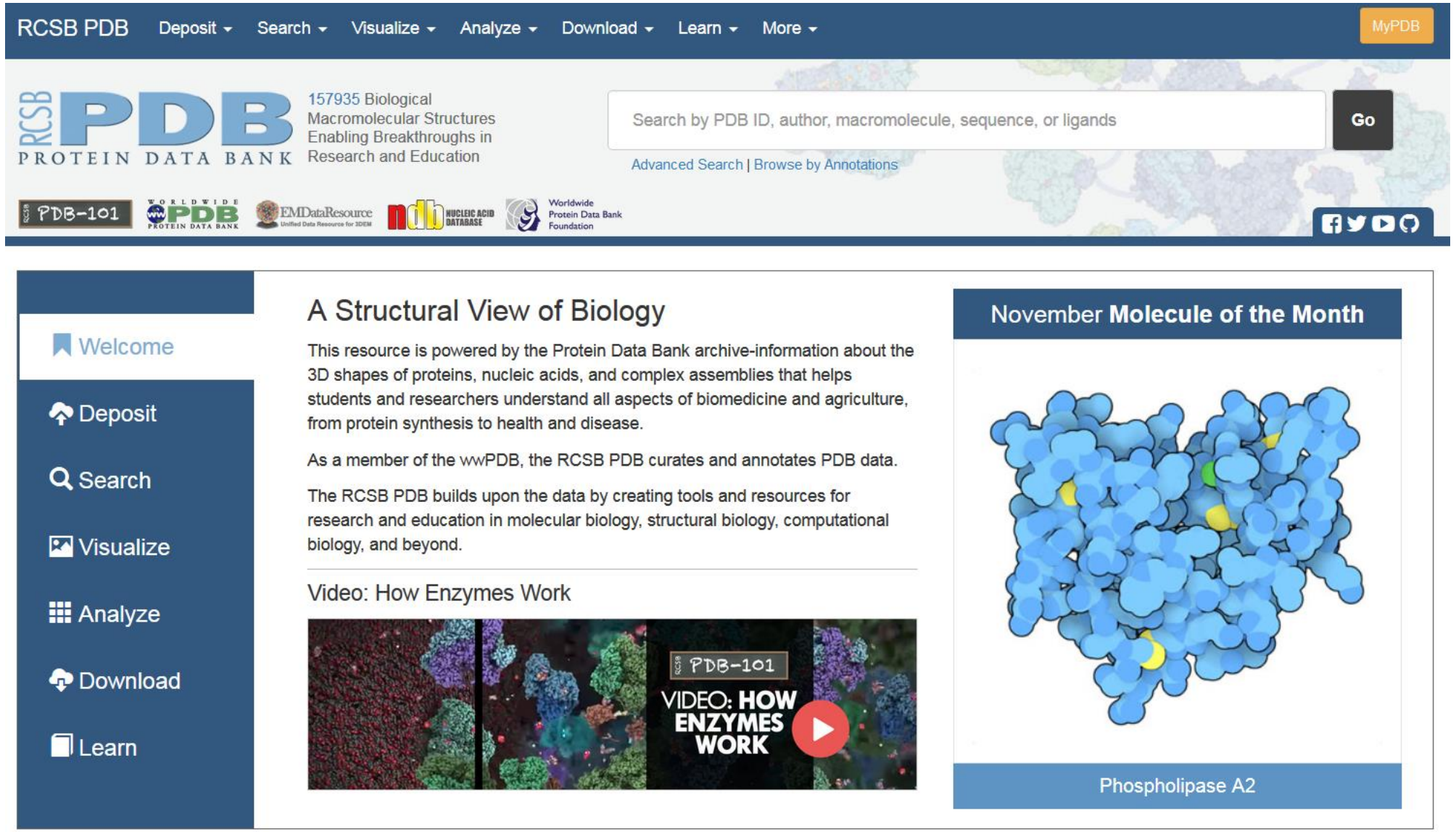
Fraction of gaps to eliminate per position 0.15 (range (0.00;1.00))

Number of sequences to consider from the Psi-Blast search 40 (range(2,50))

Your name (optional)

PROCESS RESET

Tetriary structure of proteins



Crystallin PDB search

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Manage Columns Show 100

select all *27 sequences selected* GenPept Graphics Distance tree of results Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
Chain A, Alpha-crystallin A chain [Bos taurus]	203	203	60%	3e-68	93.33%	3L1E_A
Chain A, Alpha-crystallin A chain [Bos taurus]	197	197	58%	5e-66	93.14%	3L1F_A
Chain A, Alpha-crystallin B chain [Homo sapiens]	189	189	98%	1e-61	54.49%	2KLR_A
Chain A, Alpha A crystallin [Danio rerio]	175	175	61%	3e-57	73.83%	3N3E_A
Chain C, Heat Shock Protein Beta-6 [Homo sapiens]	139	139	81%	1e-42	44.14%	5LTW_C
Chain A, Alpha-crystallin B chain [Homo sapiens]	120	120	51%	8e-36	59.55%	2N0K_A
Chain A, Human Alphab Crystallin [Homo sapiens]	119	119	57%	5e-35	54.55%	3L1G_A
Chain A, Alpha-crystallin B chain [Homo sapiens]	118	118	51%	5e-35	58.43%	6BP9_A
Chain A, ALPHA-CRYSTALLIN B CHAIN [Homo sapiens]	118	118	51%	7e-35	58.43%	2WJ7_A
Chain A, ALPHA-CRYSTALLIN B [Homo sapiens]	115	115	51%	1e-33	57.30%	2Y22_A
Chain A, Alpha-crystallin B Chain [Homo sapiens]	115	115	51%	2e-33	56.18%	2Y1Z_A
Chain A, Alpha-crystallin B Chain [Homo sapiens]	114	114	49%	2e-33	58.82%	4M5S_A
Chain A, Heat shock protein beta-1 [Homo sapiens]	115	115	86%	2e-32	41.18%	6DV5_A
Chain A, ALPHA-CRYSTALLIN B CHAIN [Homo sapiens]	111	111	47%	4e-32	58.54%	2Y1Y_A
Chain A, Alpha-crystallin B chain [Homo sapiens]	110	110	49%	6e-32	57.65%	4M5T_A
Chain A, Heat shock protein beta-1 [Homo sapiens]	105	105	50%	2e-29	55.68%	2N3J_A
Chain A, HEAT SHOCK PROTEIN BETA-6 [Rattus norvegicus]	103	103	47%	8e-29	52.44%	2WJ5_A
Chain A, Heat Shock Protein Beta-1 [Homo sapiens]	101	101	49%	3e-28	55.29%	4MJH_A
Chain A, Heat shock protein beta-1 [Homo sapiens]	101	101	49%	4e-28	55.29%	6GJH_A
Chain A, Heat Shock Protein Beta-6 [Homo sapiens]	100	100	46%	9e-28	50.62%	4JUS_A
Chain A, Heat Shock Protein Beta-6 [Homo sapiens]	99.0	99.0	46%	6e-27	49.38%	4JUT_A
Chain A, Heat Shock Protein Beta-6 [Homo sapiens]	95.5	95.5	41%	5e-26	54.93%	5LUM_A
Chain D, Heat shock protein beta-2 [Homo sapiens]	97.4	97.4	76%	2e-25	40.30%	6F2R_D

PDB 3L1E

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB

RCSB PDB PROTEIN DATA BANK 157935 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

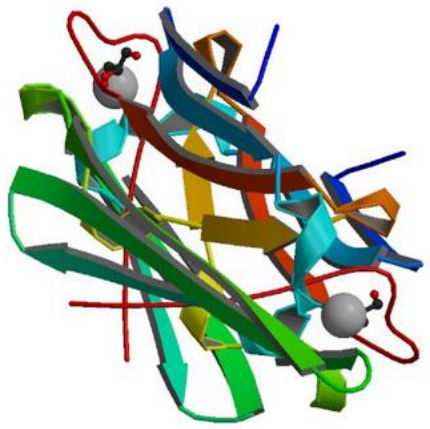
Search by PDB ID, author, macromolecule, sequence, or ligands Go

Advanced Search | Browse by Annotations

PDB-101 wwPDB EMDataResource NDB NUCLEIC ACID DATABASE Worldwide Protein Data Bank Foundation

Structure Summary | 3D View | Annotations | Sequence | Sequence Similarity | Structure Similarity | Experiment

Biological Assembly 1



3D View: Structure | Electron Density | Ligand Interaction

Standalone Viewers
Protein Workshop | Ligand Explorer

Display Files Download Files

3L1E

Bovine AlphaA crystallin Zinc Bound

DOI: 10.2210/pdb3L1E/pdb

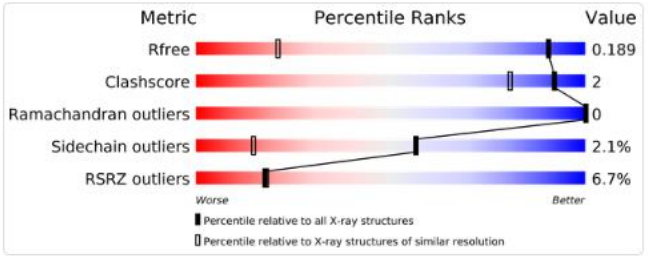
Classification: CHAPERONE
Organism(s): Bos taurus
Expression System: Escherichia coli

Deposited: 2009-12-11 **Released:** 2010-05-12
Deposition Author(s): Laganowsky, A., Sawaya, M.R., Cascio, D., Eisenberg, D.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 1.15 Å
R-Value Free: 0.193
R-Value Work: 0.164

wwPDB Validation 3D Report | Full Report



PDB 2KLR

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB

RCSB PDB PROTEIN DATA BANK

157935 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Search by PDB ID, author, macromolecule, sequence, or ligands Go

Advanced Search | Browse by Annotations

PDB-101 wwPDB EMDataResource NDB Worldwide Protein Data Bank Foundation

Structure Summary | 3D View | Annotations | Sequence | Sequence Similarity | Structure Similarity | Experiment

NMR Ensemble

3D View: Structure

Standalone Viewers
Protein Workshop | Ligand Explorer

Display Files Download Files

2KLR

Solid-state NMR structure of the alpha-crystallin domain in alphaB-crystallin oligomers

DOI: 10.2210/pdb2KLR/pdb

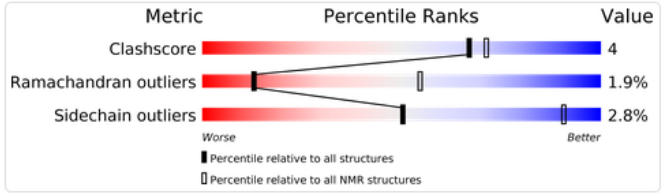
Classification: CHAPERONE
Organism(s): Homo sapiens
Expression System: Escherichia coli

Deposited: 2009-07-08 **Released:** 2010-07-07
Deposition Author(s): Jehle, S., Rajagopal, P., Markovic, S., Bardiaux, B., Kuehne, R., Higman, V.A., Klevit, R.E., van Rossum, B., Oschkinat, H.

Experimental Data Snapshot

Method: SOLID-STATE NMR
Conformers Calculated: 200
Conformers Submitted: 10
Selection Criteria: structures with the lowest energy

wwPDB Validation 3D Report | Full Report



Strict format

SUMMARY OF RECORD TYPES AND THEIR SEQUENCE

For each atomic coordinate and bibliographic entry, the file consists of records each of 80 characters. The record sequence is as follows:

HEADER	:	Date entered into Data Bank; identification code
OBSLTE	:	Identifies entries which have been replaced
COMPND	:	Name of molecule and identifying information
SOURCE	:	Species, organ, tissue, and mutant from which the molecule has been obtained, where applicable
EXPDTA	:	Experimental technique of structure determination
AUTHOR	:	Names of contributors
REVDAT	:	Revision date; identifies current modification level
SPRSDE	:	Identifies entries which have replaced others
JRNL	:	Literature citation that defines coordinate set
REMARK	:	General remarks
SEQRES	:	Residue sequence
FTNOTE	:	Footnotes relating to specific atoms or residues
HET	:	Identification of non-standard groups or residues (heterogens)
FORMUL	:	Chemical formulas of non-standard groups
HELIX	:	Identification of helical substructures
SHEET	:	Identification of sheet substructures
TURN	:	Identification of hairpin turns
SSBOND	:	Identification of disulfide bonds
SITE	:	Identification of groups comprising the various sites
CRYST1	:	Unit cell parameters, space group designation
ORIGX	:	Transformation from orthogonal Å coordinates to submitted coordinates
SCALE	:	Transformation from orthogonal Å coordinates to fractional crystallographic coordinates
MTRIX	:	Transformations expressing non-crystallographic symmetry
TVECT	:	Translation vector for infinite covalently connected structures
MODEL	:	Specification of model number for multiple structure models in a single data entry
ATOM	:	Atomic coordinate records for "standard" groups
HETATM	:	Atomic coordinate records for "non-standard" groups
SIGATM	:	Standard deviations of atomic parameters
ANISOU	:	Anisotropic temperature factors
SIGUIJ	:	Standard deviations of anisotropic temperature factors
TER	:	Chain terminator
ENDMDL	:	End-of-model flag for multiple structure models in a single data entry
CONECT	:	Connectivity records
MASTER	:	Master control record with checksums of total number of records in the file, for selected record types
END	:	End-of-entry record

ATOM Atomic coordinate records for "standard" groups
HETATM Atomic coordinate records for "non-standard" groups

Cols.	1 - 4	ATOM
or	1 - 6	HETATM
	7 - 11	Atom serial number(i)
	13 - 16	Atom name(ii)
	17	Alternate location indicator(iii)
	18 - 20	Residue name(iv,v)
	22	Chain identifier, e.g., A for hemoglobin α chain
	23 - 26	Residue seq. no.
	27	Code for insertions of residues, e.g., 66A, 66B, etc.
	31 - 38	X
	39 - 46	Y — Orthogonal Å coordinates
	47 - 54	Z
	55 - 60	Occupancy
	61 - 66	Temperature factor(vi)
	68 - 70	Footnote number

FORMAT (6A1,I5,1X,A4,A1,A3,1X,A1,I4,A1,3X,3F8.3,2F6.2,1X,I3) **tabulated**

```
TITLE     BOVINE ALPHAA CRYSTALLIN ZINC BOUND
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: ALPHA-CRYSTALLIN A CHAIN;
COMPND   3 CHAIN: A;
COMPND   4 FRAGMENT: RESIDUES 59-163;
COMPND   5 SYNONYM: ALPHA-CRYSTALLIN A CHAIN, SHORT FORM;
COMPND   6 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: BOS TAURUS;
SOURCE   3 ORGANISM_COMMON: BOVINE;
SOURCE   4 ORGANISM_TAXID: 9913;
SOURCE   5 GENE: CRYA1, CRYAA;
SOURCE   6 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   7 EXPRESSION_SYSTEM_TAXID: 562;
SOURCE   8 EXPRESSION_SYSTEM_STRAIN: BL21(DE3);
SOURCE   9 EXPRESSION_SYSTEM_VECTOR_TYPE: PLASMID;
SOURCE  10 EXPRESSION_SYSTEM_PLASMID: PET28
KEYWDS    LENS TRANSPARENCY, POLYDISPERSITY, PROTEIN AGGREGATION, CRYSTALLIN,
KEYWDS   2 EYE LENS PROTEIN, CHAPERONE
EXPDTA    X-RAY DIFFRACTION
AUTHOR    A.LAGANOWSKY,M.R.SAWAYA,D.CASCIO,D.EISENBERG
REVDAT   3   01-NOV-17 3L1E    1       REMARK
REVDAT   2   14-JUL-10 3L1E    1       JRNL
REVDAT   1   12-MAY-10 3L1E    0
JRNL        AUTH   A.LAGANOWSKY,J.L.BENESCH,M.LANDAU,L.DING,M.R.SAWAYA,
JRNL        AUTH 2 D.CASCIO,Q.HUANG,C.V.ROBINSON,J.HORWITZ,D.EISENBERG
JRNL        TITL   CRYSTAL STRUCTURES OF TRUNCATED ALPHAA AND ALPHAB
JRNL        TITL 2 CRYSTALLINS REVEAL STRUCTURAL MECHANISMS OF POLYDISPERSITY
JRNL        TITL 3 IMPORTANT FOR EYE LENS FUNCTION.
JRNL        REF    PROTEIN SCI.                  V.  19  1031 2010
JRNL        REFN                   ISSN 0961-8368
JRNL        PMID   20440841
JRNL        DOI    10.1002/PRO.380
REMARK   2
REMARK   2 RESOLUTION.    1.15 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : PHENIX 1.5_2
REMARK   3   AUTHORS     : PAUL ADAMS,PAVEL AFONINE,VINCENT CHEN,IAN
REMARK   3               : DAVIS,KRESHNA GOPAL,RALF GROSSE-KUNSTLEVE,
REMARK   3               : LI-WEI HUNG,ROBERT IMMORMINO,TOM IOERGER,
REMARK   3               : AIRLIE MCCOY,ERIK MCKEE,NIGEL MORIARTY,
REMARK   3               : REETAL PAI,RANDY READ,JANE RICHARDSON,
REMARK   3               : DAVID RICHARDSON,TOD ROMO,JIM SACCHETTINI,
REMARK   3               : NICHOLAS SAUTER,JACOB SMITH,LAURENT
REMARK   3               : STORONI,TOM TERWILLIGER,PETER ZWART
```

...

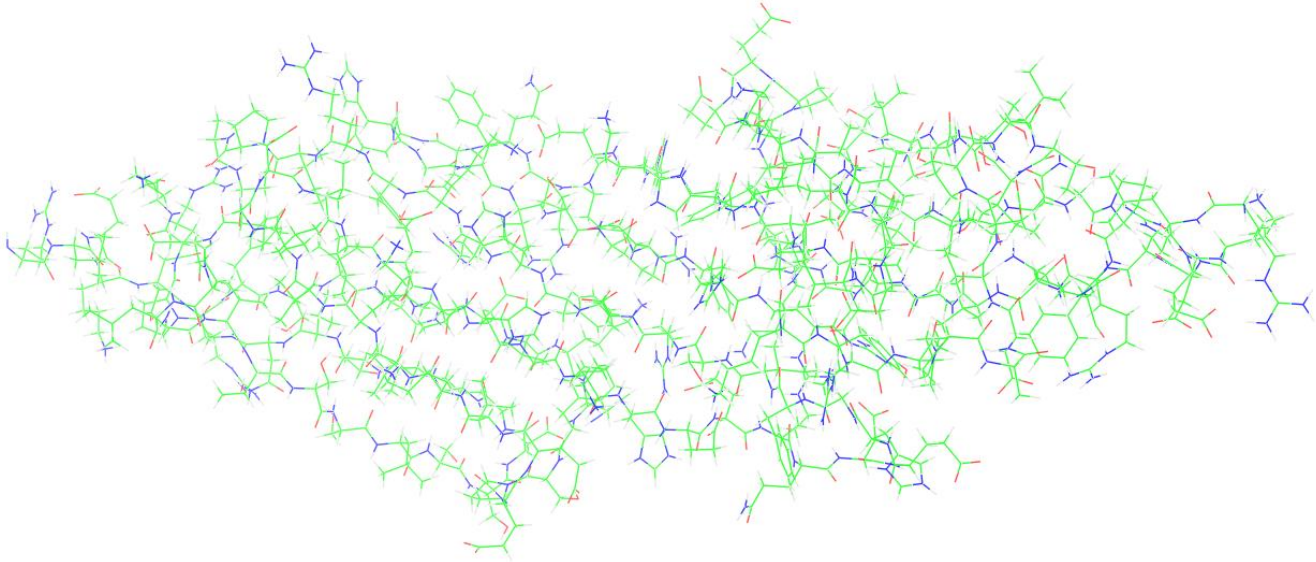
```
SEQRES   1 A  106  GLY SER GLY ILE SER GLU VAL ARG SER ASP ARG ASP LYS
SEQRES   2 A  106  PHE VAL ILE PHE LEU ASP VAL LYS HIS PHE SER PRO GLU
SEQRES   3 A  106  ASP LEU THR VAL LYS VAL GLN GLU ASP PHE VAL GLU ILE
SEQRES   4 A  106  HIS GLY LYS HIS ASN GLU ARG GLN ASP ASP HIS GLY TYR
SEQRES   5 A  106  ILE SER ARG GLU PHE HIS ARG ARG TYR ARG LEU PRO SER
SEQRES   6 A  106  ASN VAL ASP GLN SER ALA LEU SER CYS SER LEU SER ALA
SEQRES   7 A  106  ASP GLY MET LEU THR PHE SER GLY PRO LYS ILE PRO SER
SEQRES   8 A  106  GLY VAL ASP ALA GLY HIS SER GLU ARG ALA ILE PRO VAL
SEQRES   9 A  106  SER ARG
HELIX    1   1 SER A   81  GLU A   83  5                                   3
SHEET    1   A 4 SER A  62  SER A  66  0
SHEET    2   A 4 LYS A  70  ASP A  76 -1  O  VAL A  72   N  ARG A  65
SHEET    3   A 4 MET A 138  PRO A 144 -1  O  LEU A 139   N  LEU A  75
SHEET    4   A 4 SER A 130  LEU A 133 -1  N  SER A 132   O  THR A 140
SHEET    1   B 3 LEU A  85  GLN A  90  0
SHEET    2   B 3 PHE A  93  GLN A 104 -1  O  GLU A  95   N  LYS A  88
SHEET    3   B 3 GLY A 108  ARG A 119 -1  O  TYR A 118   N  VAL A  94
LINK         OE2 GLU A 102                ZN    ZN A   1     1555   1555  1.96
LINK         NE2 HIS A 100                ZN    ZN A   1     1555   1555  1.99
SITE     1 AC1  4 HIS A 100  GLU A 102  HIS A 107  HIS A 154
SITE     1 AC2  8 HOH A   5  HOH A   9  HOH A  29  GLN A 104
SITE     2 AC2  8 ASP A 105  ARG A 116  SER A 122  HOH A 211
CRYST1   56.215   56.215   68.657  90.00  90.00  90.00 P 41 21 2     8
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.017789  0.000000  0.000000        0.00000
SCALE2      0.000000  0.017789  0.000000        0.00000
SCALE3      0.000000  0.000000  0.014565        0.00000
ATOM      1  N   SER A  59      17.064  24.661  22.613  1.00 24.05           N
ATOM      2  CA  SER A  59      16.108  25.532  23.283  1.00 23.26           C
ATOM      3  C   SER A  59      15.419  24.799  24.428  1.00 22.63           C
ATOM      4  O   SER A  59      14.187  24.687  24.468  1.00 22.68           O
ATOM      5  CB  SER A  59      15.073  26.052  22.289  1.00 22.64           C
ATOM      6  OG  SER A  59      14.265  27.041  22.892  1.00 21.53           O
ATOM      7  N   GLY A  60      16.224  24.289  25.353  1.00 21.92           N
ATOM      8  CA  GLY A  60      15.715  23.602  26.523  1.00 20.45           C
ATOM      9  C   GLY A  60      15.363  22.153  26.248  1.00 18.99           C
ATOM     10  O   GLY A  60      15.869  21.539  25.307  1.00 18.83           O
ATOM     11  N   ILE A  61      14.480  21.605  27.074  1.00 18.06           N
ATOM     12  CA  ILE A  61      14.105  20.203  26.966  1.00 17.28           C
ATOM     13  C   ILE A  61      13.407  19.881  25.641  1.00 13.75           C
ATOM     14  O   ILE A  61      13.715  18.874  25.005  1.00 12.94           O
ATOM     15  CB  ILE A  61      13.183  19.778  28.131  1.00 20.26           C
ATOM     16  CG1 ILE A  61      12.871  18.289  28.048  1.00 21.30           C
ATOM     17  CG2 ILE A  61      11.870  20.528  28.088  1.00 21.55           C
ATOM     18  CD1 ILE A  61      11.382  17.970  27.968  1.00 22.55           C
```

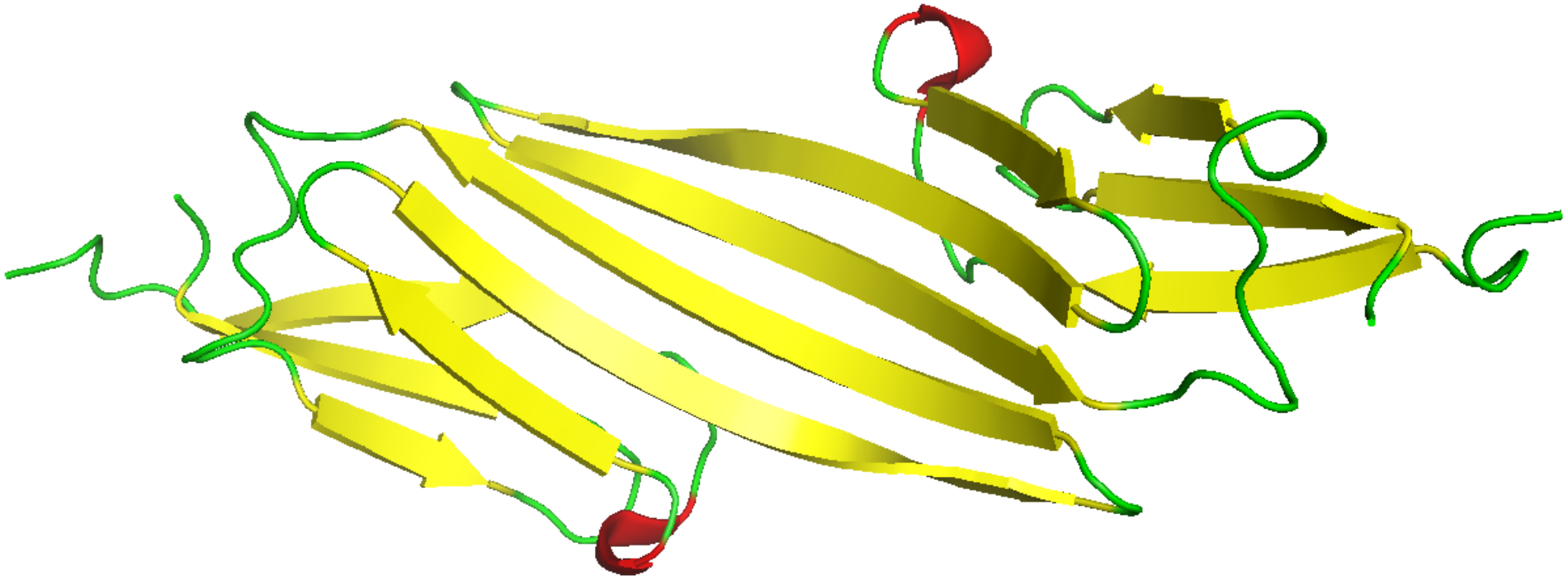
...

Visualizing PDB structures

- PyMOL
- Chimera
- VMD
- Maestro

PDB 2KLR,PyMOL





Classification of protein structures

Comparison of SCOP and CATH Hierarchies

SCOP	CATH
class	class
	architecture
fold	topology
	homologous superfamily
superfamily	↓
family	sequence family
domain	domain

all α, **all** β, α/β, α+β, α&β, **membrane, small,…**

CATH more directed toward structural classification,
SCOP pays more attention to evolutionary relationships

The use of protein structures

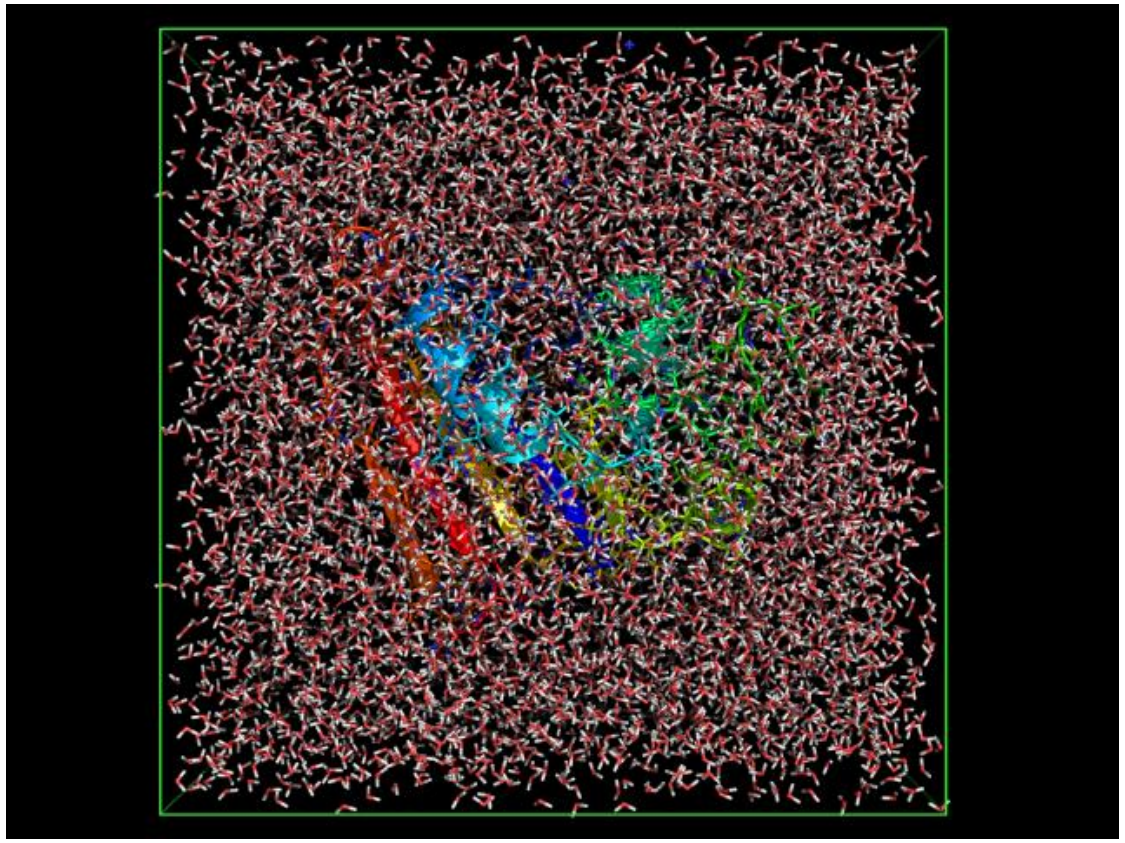
Understanding protein function / enzymatic catalysis

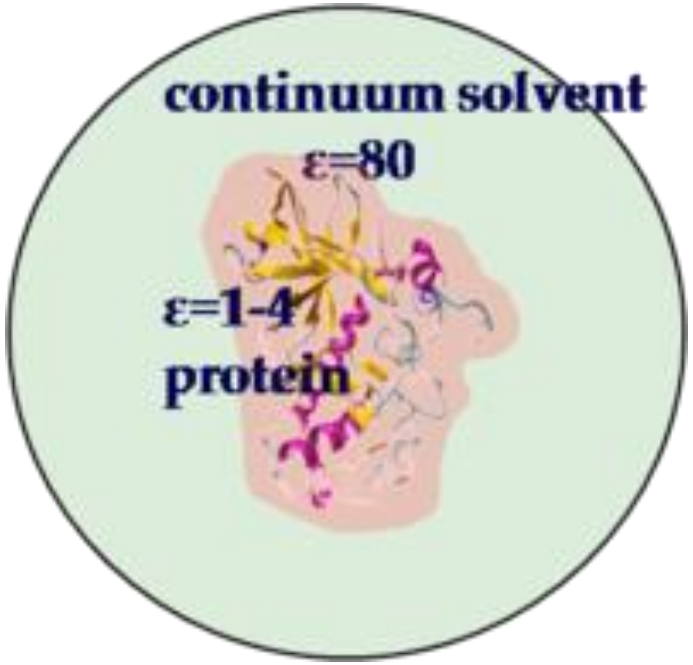
Understanding protein-ligand interaction / drug design

Molecular mechanics

- Classical physics model
- Forcefields describing system, internal coordinates
- Quantum chemistry is „expensive”
- Hybrid simulations in the case of bond formation
- One proteins conformation – one energy value
- Local minimum search, energy minimization

Solvation





Explicit Implicit

accounts for the hydrophobic effect

Forcefields

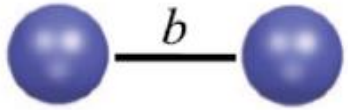
Force fields: united / all atom: atom types; hydrogens

AMBER

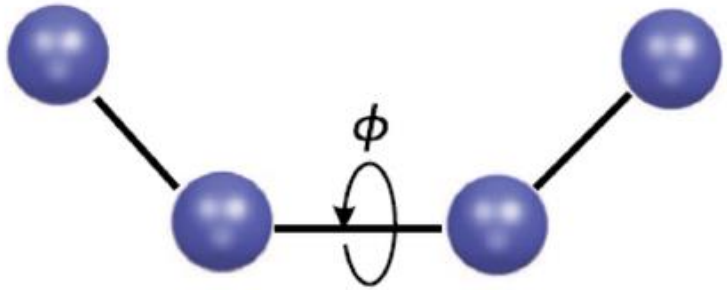
CVFF

CHARMM

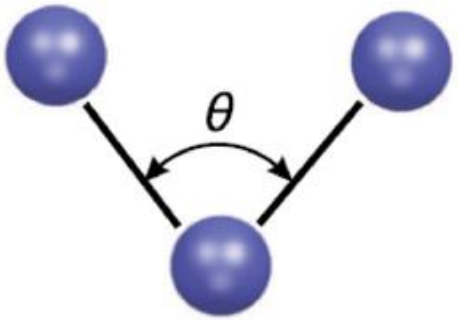
GROMOS



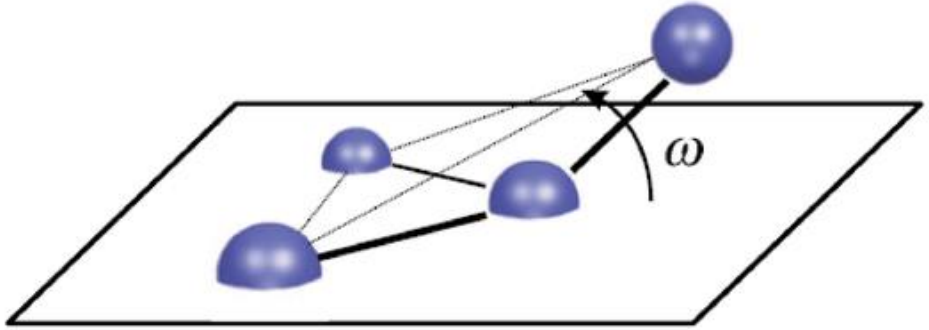
Bond stretching



Proper dihedral torsion



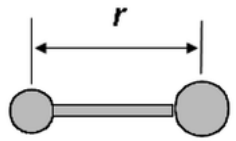
Angle bending



Improper dihedral torsion

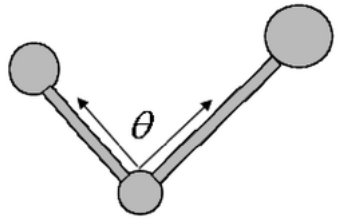
Physical model for the AMBER force field

$$V_{total} = V_{bond} + V_{angle} + V_{torsion} + V_{non-bond}$$



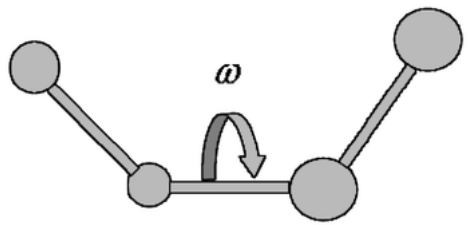
$$V_{bond} = k_{bond}(r - r_0)^2$$

(a)



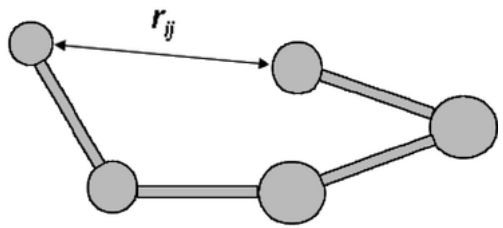
$$V_{angle} = k_{angle}(\theta - \theta_0)^2$$

(b)



$$V_{torsion} = \frac{1}{2}k_{torsion}\{1 + \cos(n\omega - \omega_0)\}$$

(c)



$$V_{non-bond} = \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}}$$

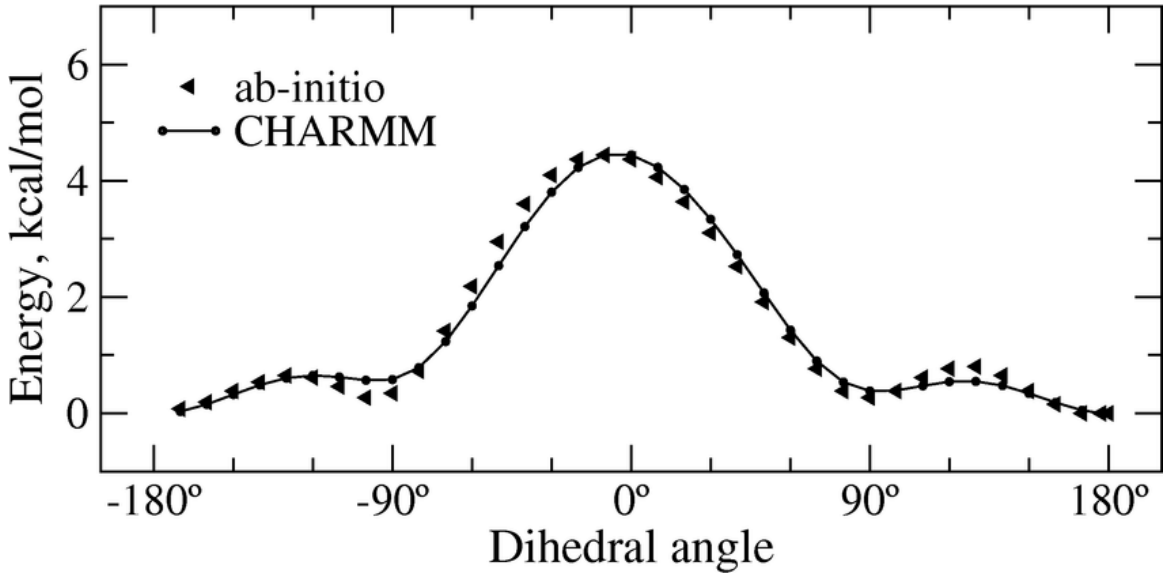
(d)

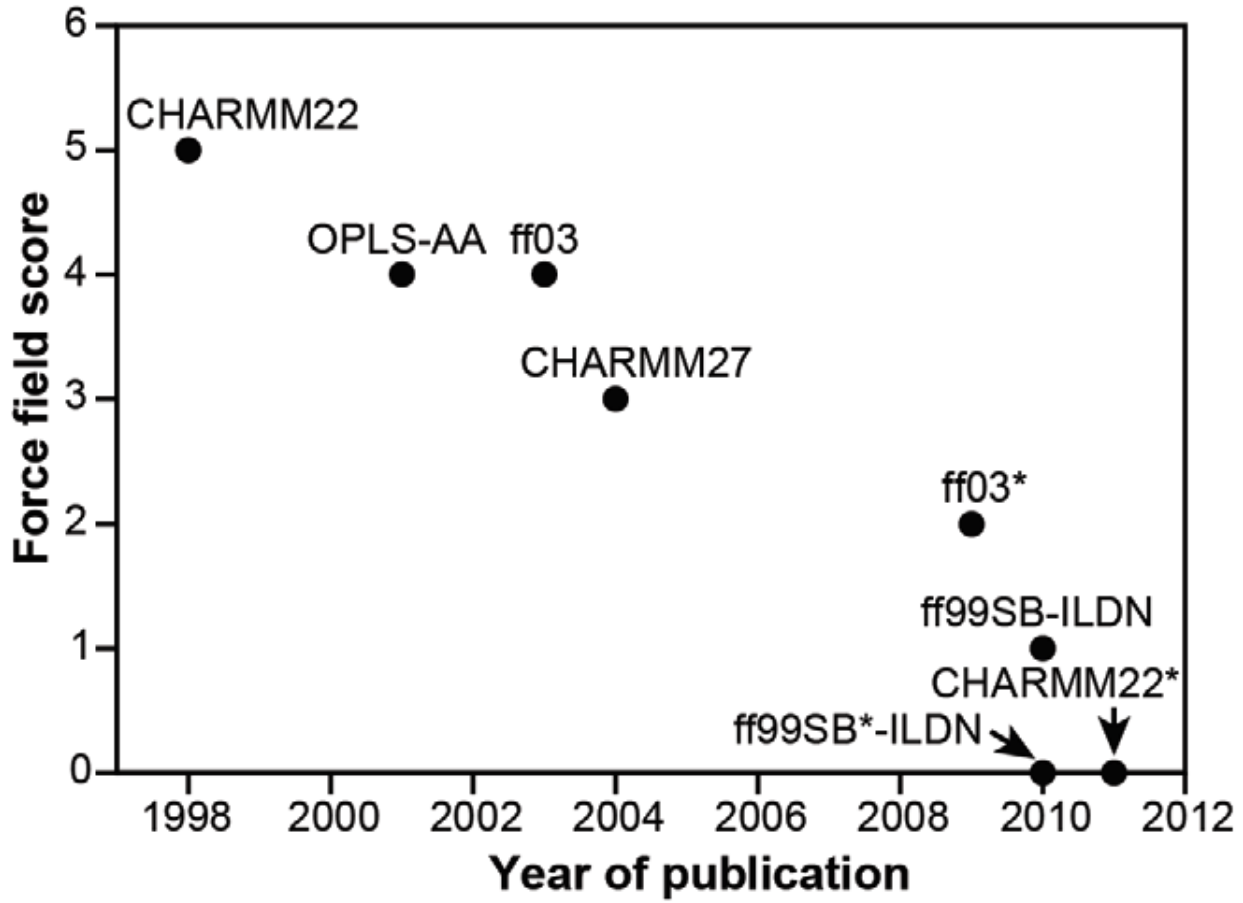
$$V(r^N) = \sum_{\text{bonds}} k_b(l - l_0)^2 + \sum_{\text{angles}} k_a(\theta - \theta_0)^2$$

$$+ \sum_{\text{torsions}} \sum_{n} \frac{1}{2} V_n[1 + \cos(n\omega - \gamma)] + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} f_{ij} \left\{ \epsilon_{ij} \left[\left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{0ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\}$$

Forcefield parametrization, accuracy

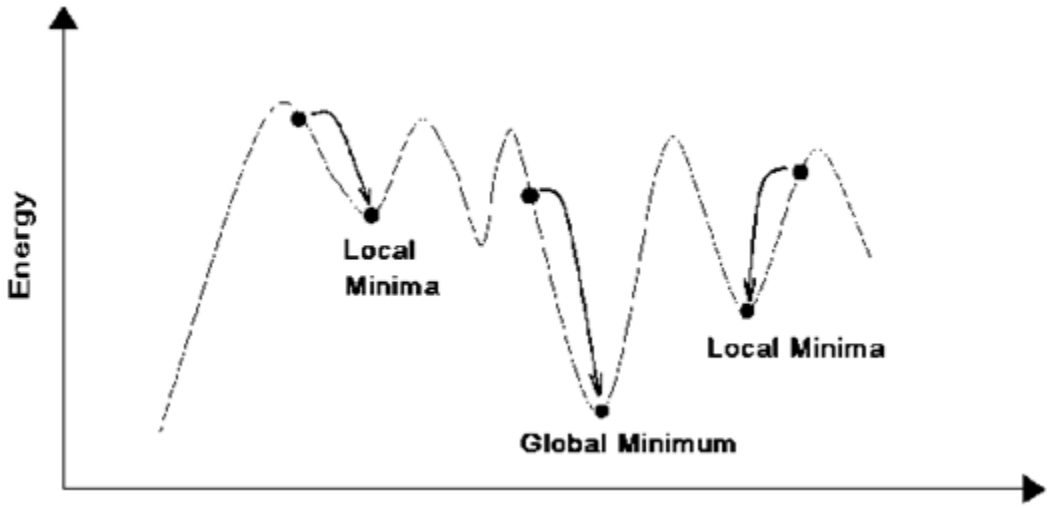
empirical parameters
ab initio calculations



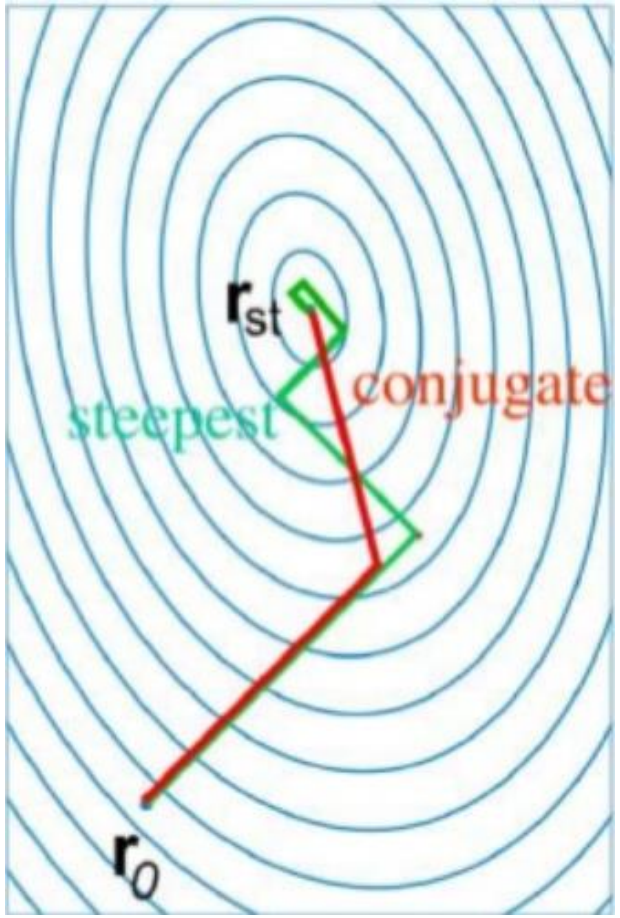


Local energy minimum search

energy minimization



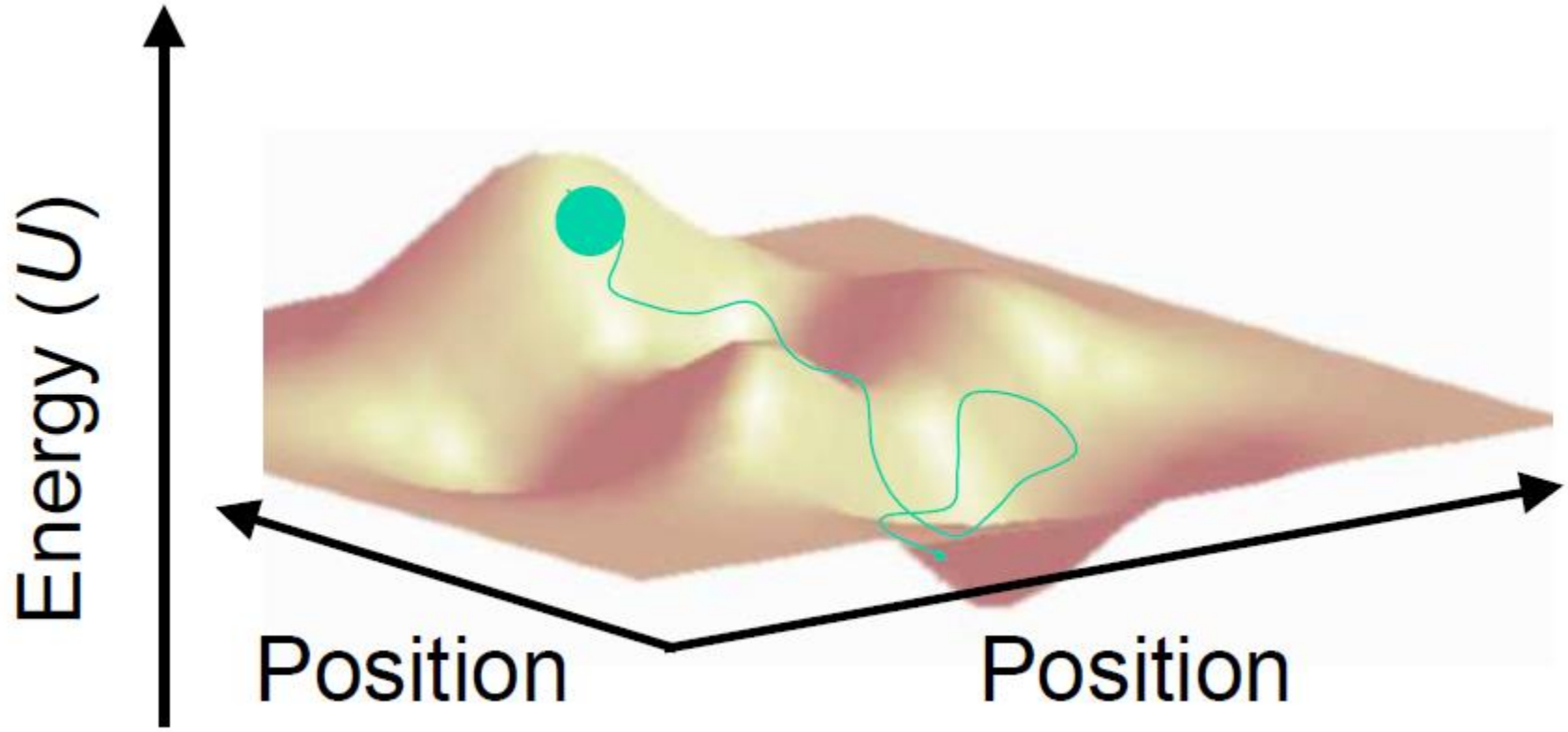
The global minimum is relevant



Molecular dynamics

- Equation of motion $\vec{F}$=m $\frac{d\vec{v}}{dt}$, where $\vec{v}$=$\frac{d\vec{x}}{dt}$
- Initial velocity distribution, simulation temperature
- Simulated annealing
- dt timestep ~ fsec range
- Numerical integration, Verlet algorithm
- Sampling

Dicovering the potential energy surface



Use of MD simulations

- Global minimum search
  - protein folding
  - homology model refinement
- Free Energy Perturbation
  - effect of mutations
  - ligand binding

Popular MD programs

- GROMOS
- NAMD
- AMBER
- DESMOND

GPU acceleration

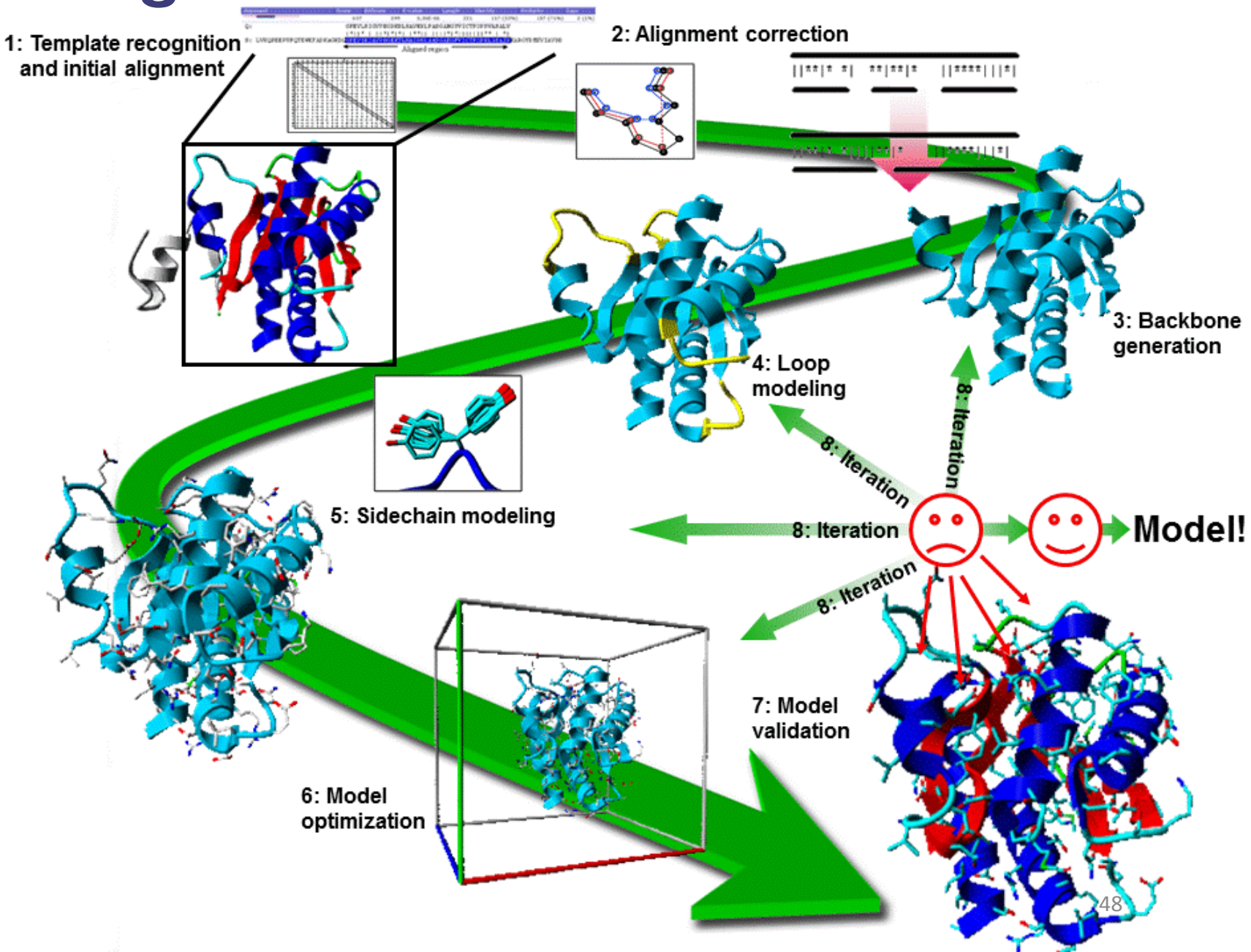
microsecond range available on a desktop

Folding happens in the micro-millisecond range

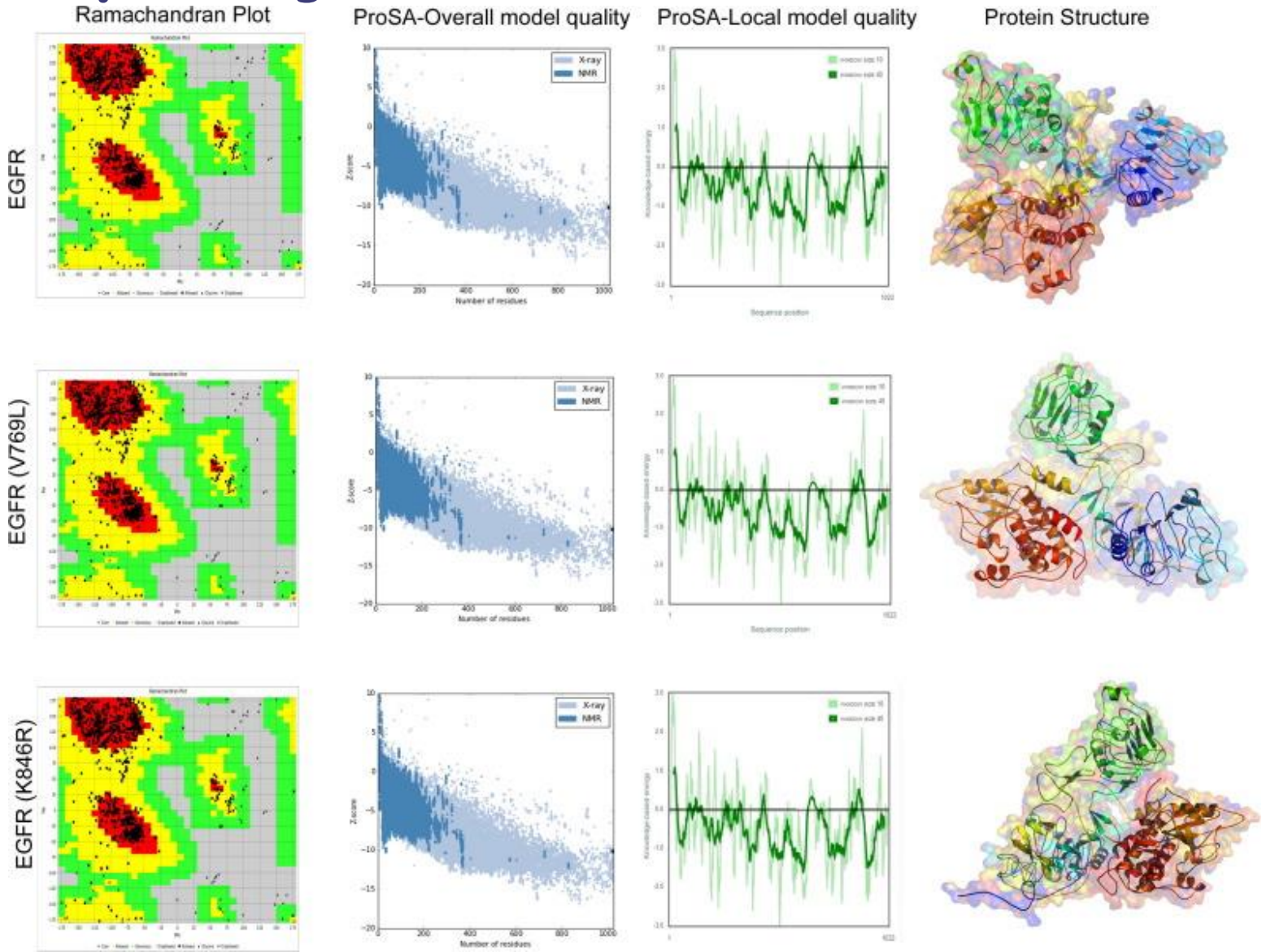
Homology modelling

Popular programs

MODELLER
iTASSER
Swiss Model



Model quality assessment



Critical Assessment of protein Structure Prediction

- CASP12 in numbers

Number of groups registered	**192**
including: *expert groups*	*112*
prediction servers	*80*
Number of regular targets released	**82**
including *all-group (human) targets*	*56*
Targets canceled and not re-released for all/manual prediction	**11 / 11**
Number of refinement targets released	**42**
Number of assisted prediction targets released	**14**

- I-TASSER

I-TASSER

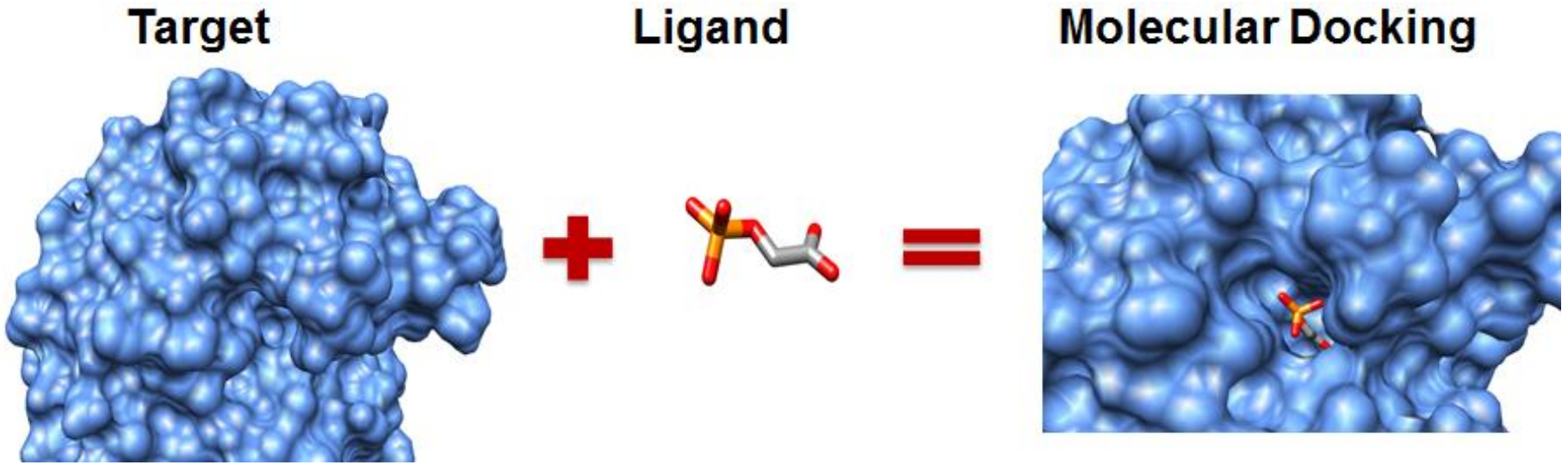
Protein Structure & Function Predictions

(The server completed predictions for 379638 proteins submitted by 92066 users from 136 countries)
(The template library was updated on 2018/02/11)

I-TASSER (Iterative Threading ASSEmbly Refinement) is a hierarchical approach to protein structure and function prediction. It first identifies structural templates from the PDB by multiple threading approach LOMETS, with full-length atomic models constructed by iterative template fragment assembly simulations. Function insights of the target are then derived by threading the 3D models through protein function database BioLiP. I-TASSER (as 'Zhang-Server') was ranked as the No 1 server for protein structure prediction in recent community-wide CASP7, CASP8, CASP9, CASP10, CASP11, and CASP12 experiments. It was also ranked as the best for function prediction in CASP9. The server is in active development with the goal to provide the most accurate structural and function predictions using state-of-the-art algorithms. Please report problems and questions at I-TASSER message board and our developers will answer the questions. (>> More about the server ...)

- http://zhanglab.ccmb.med.umich.edu/I-TASSER/

Protein-ligand interactions



vHTS, multi million compound libraries; scoring functions
Pose prediction; MM
Lead optimization; MD FEP

Scoring functions

$$\Delta G_{\text{bind}} = C_{\text{lipo-lipo}} \sum f(r_{\text{lr}}) + C_{\text{hbond-neut-neut}} \sum g(\Delta r)\, h(\Delta\alpha) + C_{\text{hbond-neut-charged}} \sum g(\Delta r)\, h(\Delta\alpha) + C_{\text{hbond-charged-charged}} \sum g(\Delta r)\, h(\Delta\alpha) + C_{\text{max-metal-ion}} \sum f(r_{\text{lm}}) + C_{\text{rotb}} H_{\text{rotb}} + C_{\text{polar-phob}} V_{\text{polar-phob}} + C_{\text{coul}} E_{\text{coul}} + C_{\text{vdW}} E_{\text{vdW}} + \text{solvation terms}$$

Popular programs / scoring functions
Autodock, GOLD, Glide

Receptor flexibility

- Rigid receptor model can be used on some targets
- Ensemble docking
- Induced Fit Docking