Компьютерный анализ белковой последовательности

Анализируют только аминокислотную последовательность белка, пренебрегают взаимодействием между боковыми цепями аминокислот.

Что можно делать:

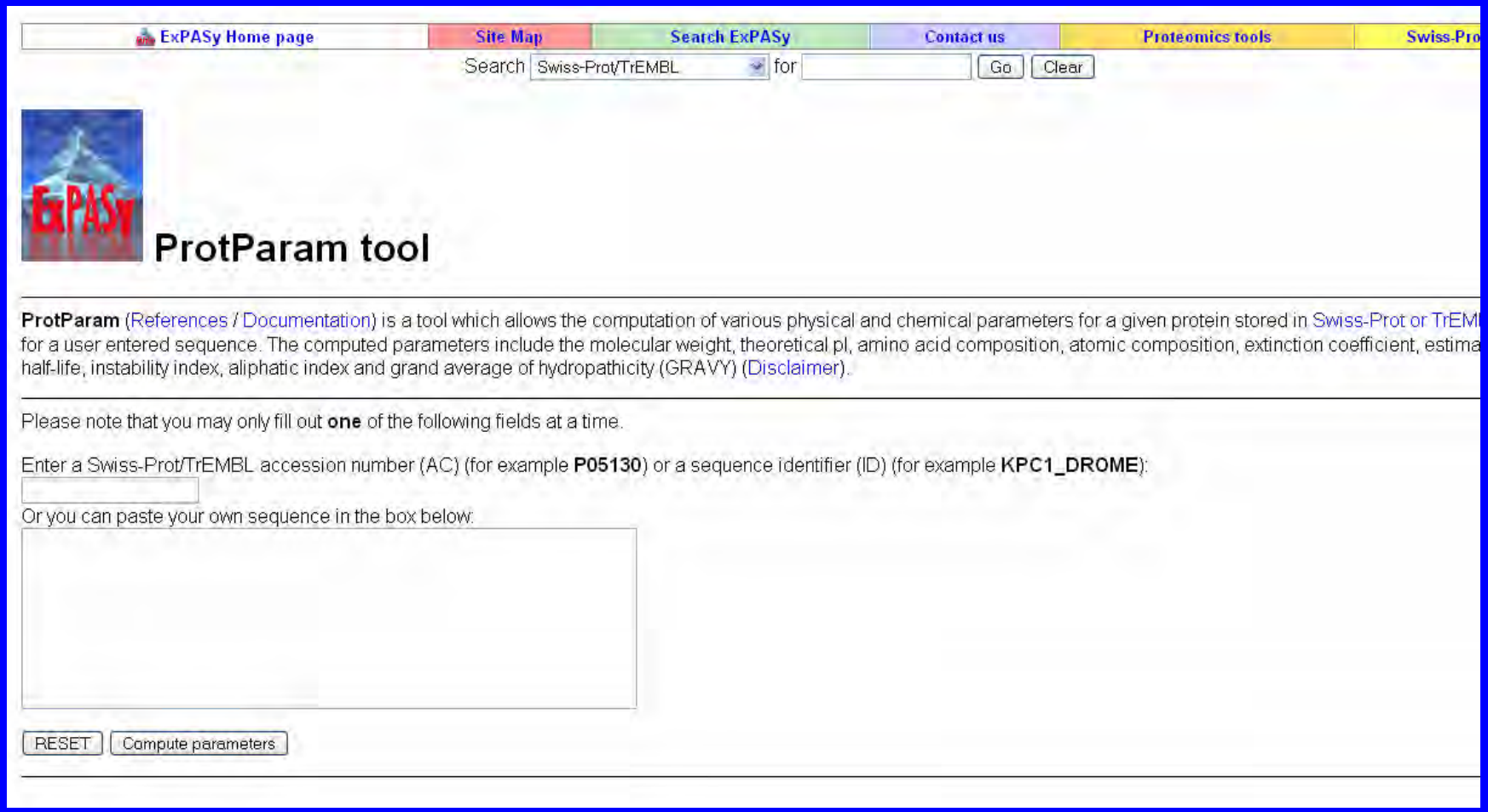
- ✓ Вычисление физико-химических параметров белка
- ✓ Предсказание продуктов расщепления протеазами
- ✓ Гидрофобные, гидрофильные участки: например, трансмембранные сегменты
- ✓ Пост-трансляционные модификации
- ✓ Функциональные домены, принадлежность к функциональным семействам

Компьютерный анализ белковой последовательности

… и где это можно делать:

- ✓ The ExPASy server – протеомика *http://www.expasy.ch/tools/#primary*
- ✓ The Swiss EMBnet – coiled-coil участки, выравнивания и др. биоинф. анализ *http://www.ch.embnet.org*
- ✓ The **CBS** Prediction Servers – локализация, пост-трансляционные модификации… http://www.cbs.dtu.dk/services

Программы для предсказания физико-химических параметров белка: ProtParam



ProtParam

ProtParam

Selection of endpoints on the sequence

EGFR_HUMAN (P00533)

Epidermal growth factor receptor precursor (EC 2.7.10.1) (Receptortyrosine-protein kinase ErbB-1).
Homo sapiens (Human).

Please select one of the following features by clicking on a pair of endpoints, and the computation will be carried out for the corresponding sequence fragment.
Note: Only the features corresponding to subsequences of at least 5 residues are highlighted.

FT	SIGNAL	1-24	
FT	**CHAIN**	**25-1210**	**Epidermal growth factor receptor.**
FT	TOPO_DOM	25-645	Extracellular (Potential).
FT	TRANSMEM	646-668	Potential.
FT	TOPO_DOM	669-1210	Cytoplasmic (Potential).
FT	REPEAT	75-300	Approximate.
FT	REPEAT	390-600	Approximate.
FT	DOMAIN	712-979	Protein kinase.
FT	NP_BIND	718-726	ATP (By similarity).
FT	COMPBIAS	1025-1071	Ser-rich.
FT	STRAND	40-43	
FT	HELIX	44-55	
FT	STRA		
FT	STRA		
FT	HELI		
FT	STRA		

FT	STRAND	850-853
FT	HELIX	858-861
FT	TURN	862-865
FT	TURN	878-880
FT	HELIX	883-888
FT	HELIX	893-908
FT	TURN	914-917
FT	HELIX	920-922
FT	HELIX	923-929
FT	HELIX	941-950
FT	HELIX	955-957
FT	HELIX	961-973
FT	HELIX	975-978
FT	TURN	982-986
FT	HELIX	996-1002
FT	HELIX	1013-1016

Or, if you wish to select a different sequence fragment (at least 5 amino acids long), you can enter the desired endpoints on the sequence here (by default, the computation will b
carried out for the complete sequence).

N-terminal: 5
C-terminal: 25

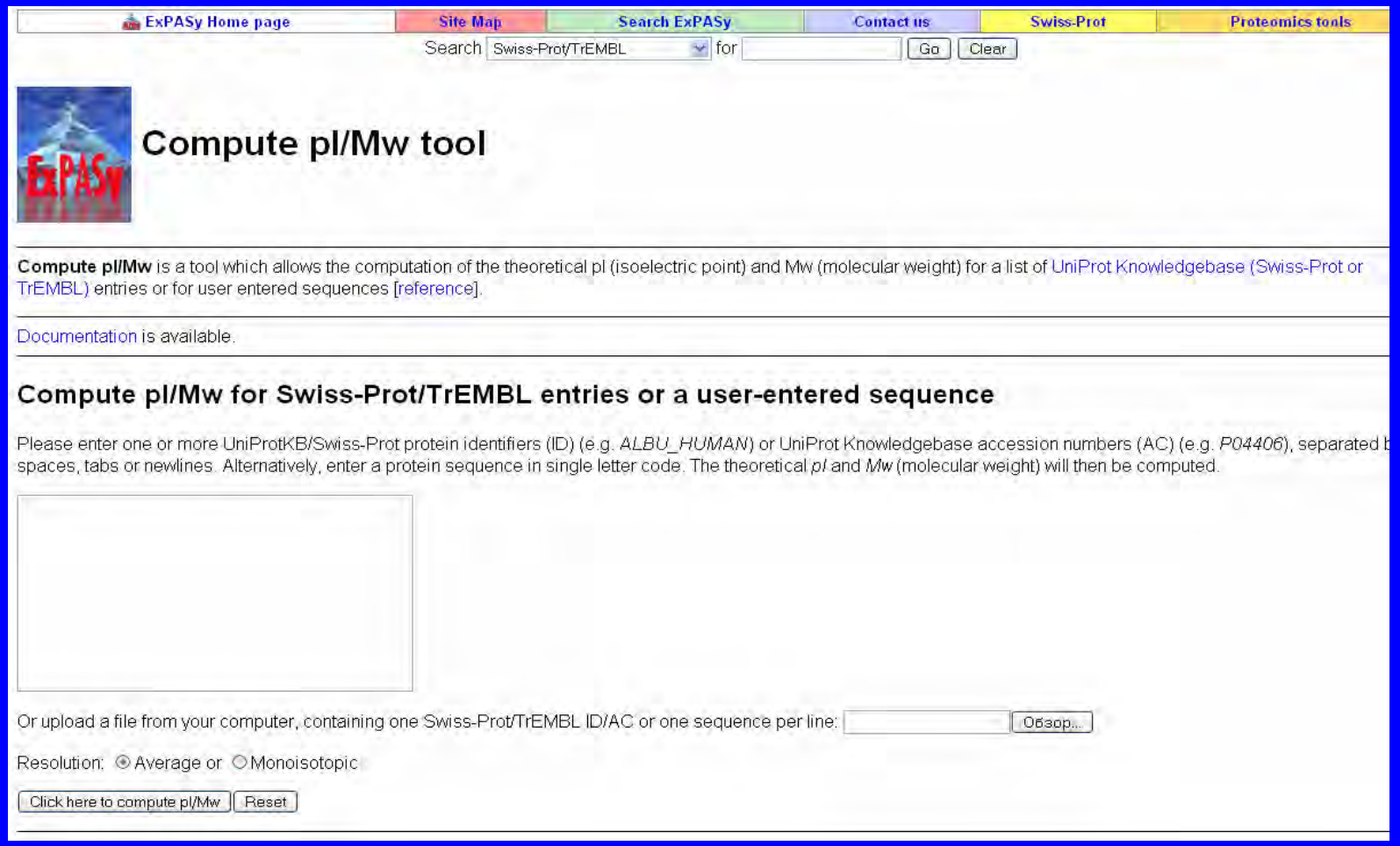
The sequence EGFR_HUMAN consists of 1210 amino acids.

RESET SUBMIT

ProtParam

- ✓ Molecular weight (не учитывает пост-трансляционных модификаций)
- ✓ Аминокислотный состав
- ✓ Теоретическая pI
- ✓ Extinction coefficients (280 nm) (не учитывает пространственных взаимодействий аминокислот)
- ✓ Instability (менее 40 – хорошо) – нестабильность в эксперименте (test tube, статистика дипептидов)
- ✓ Half-life (yeast in vivo, mammalian reticulocytes in vitro, Escherichia coli in vivo; N-terminal rule: время полужизни определяется N-терминальной аминокислотой)
- ✓ Алифатический индекс
- ✓ Grand average of hydropathicity (GRAVY) гидрофильность – (-), гидрофобность – (+)

Compute pI/Mw



Выбирается участок белка (или весь белок), для него вычисляются теоретическая pI и молекулярный вес

Простейшие программы по вычислению параметров: PeptideMass

Можно учитывать или не учитывать пост-трансляционные модификации для белков из Swiss-Prot, а также полиморфизмы, AS изоформы и конфликты

PeptideMass - output

PeptideMass

The entered protein is: P00533

The selected enzyme is: Trypsin

Maximum number of missed cleavages (MC): 0

All cysteines in reduced form.

Methionines have not been oxidized.

Displaying peptides with a mass bigger than 500 Dalton.

Using monoisotopic masses of the occurring amino acid residues and giving peptide masses as $[M+H]^+$.

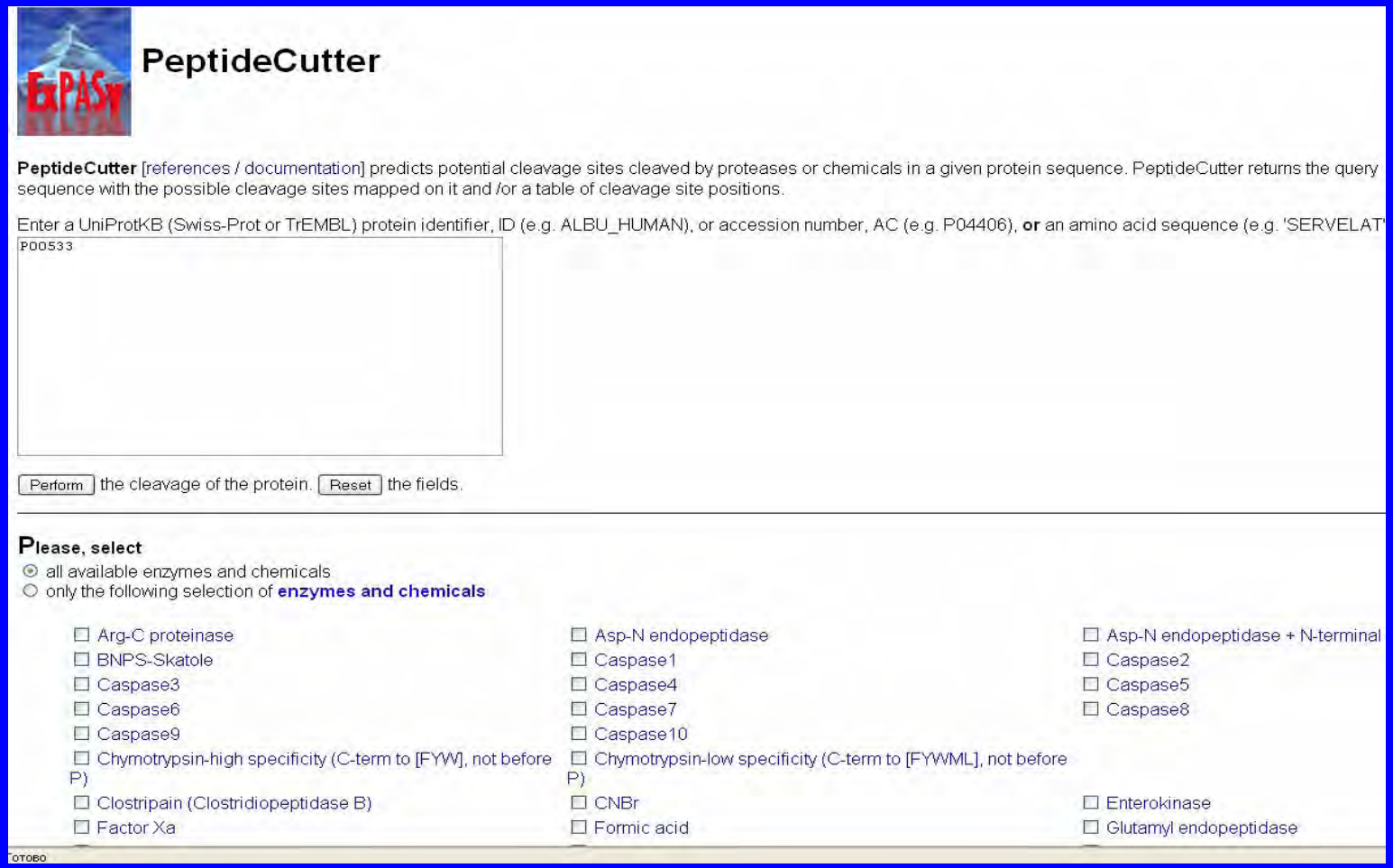
You have selected EGFR_HUMAN (P00533) from UniProtKB/Swiss-Prot:

Epidermal growth factor receptor precursor (EC 2.7.10.1) (Recepto tyrosine-protein kinase ErbB-1)
Signal in positions 1-24 has been removed.

- Chain Epidermal growth factor receptor at positions 25 - 1210 [Theoretical pI: 6.17 / Mw (average mass): 132012.67 / Mw (monoisotopic mass): 131927.02]

mass	position	#MC	modifications		peptide sequence
4534.2162	890-929	0			IYTHQSDVWSYGVTVWELMT FGSKPYDGIPASEISSILEK
4351.9737	1122-1160	0			DPHYQDPHSTAVGNPEYLNT VQPTCVNSTFDSPAHWAQK
3607.5726	1000-1031	0	PHOS: 1026	3687.5390	ALMDEEDMDDVVDADEYLIP QQGFFSSPSTSR
3398.6161	1069-1099	0	PHOS: 1070, 1071	3558.5488	YSSDPTGALTEDSIDDTFLP VPEYINQSVPK
3372.4111	610-642	0			YADAGHVCHLCHPNCTYGCT GPGLEGCPTNGPK
3052.3090	548-574	0			EFVENSECIQCHPECLPQAM NITCTGR
3010.6178	777-803	0			LLGICLTSTVQLITQLMPFG CLLDYVR
2738.6438	643-669	0			IPSIATGMVGALLLLLVVAL GIGLFMR
2703.1847	166-189	0			DIVSSDFLSNMSMDFQNHLG SCQK
2627.1826	262-284	0			DTCPPLMLYNPTTYQMDVNP EGK
2400.0812	109-129	0			GNMYYENSYALAVLSNYDAN K

PeptideCutter



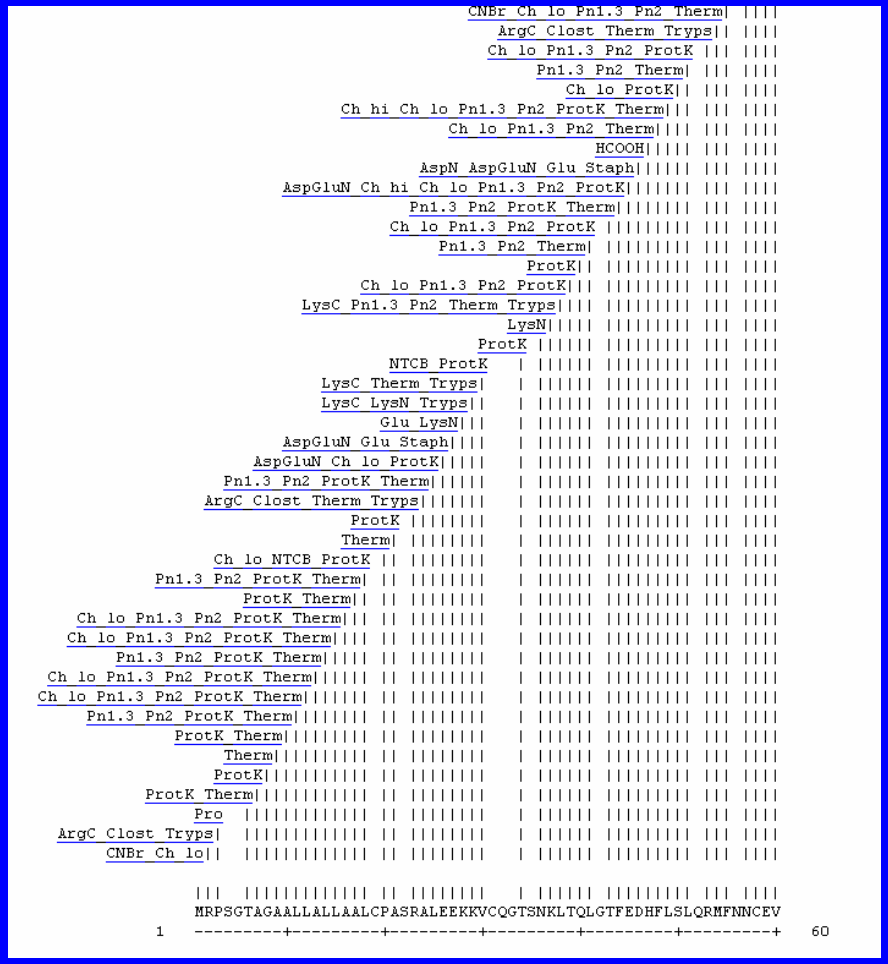
Для трипсина и хемотрипсина можно выбрать другую модель, в которой будет посчитана вероятность расщепления по каждому остатку

PeptideCutter - output

These enzymes cleave the sequence:

Name of enzyme	No. of cleavages	Positions of cleavage sites
Arg-C proteinase	60	2 23 53 72 98 108 138 149 165 222 224 244 252 255 297 309 324 334 377 414 427 429 451 494 521 527 531 533 547 574 669 670 671 675 677 68 681 686 705 748 776 803 831 832 836 841 889 932 958 962 973 977 986 999 1031 1052 1068 1100 1121 1199
Asp-N endopeptidase	61	45 74 125 165 170 178 190 229 246 255 261 277 302 313 320 346 367 378 387 392 415 457 459 521 536 576 586 611 760 769 799 806 829 836 8 895 915 941 953 955 973 983 993 1002 1005 1007 1008 1011 1013 1050 1062 1071 1079 1082 1083 1121 1126 1151 1167 1170 1174
Asp-N endopeptidase + N-terminal Glu	138	25 26 44 45 58 65 74 83 96 101 113 125 133 141 159 165 170 178 190 203 204 229 244 246 255 256 261 277 281 302 313 316 318 319 320 329 3 346 367 378 387 390 392 399 411 415 420 423 454 457 459 495 512 518 521 533 536 542 544 547 550 553 560 576 586 601 611 633 684 686 68 708 710 733 735 745 748 757 760 761 769 799 803 806 828 829 836 854 864 865 867 871 883 895 905 915 921 927 930 941 953 955 962 966 97 984 993 1002 1003 1004 1005 1007 1008 1011 1013 1014 1050 1061 1062 1071 1078 1079 1082 1083 1090 1121 1126 1136 1151 1167 1170 117 1179 1192 1195 1205
BNPS-Skatole	13	164 200 410 477 516 608 731 817 880 898 905 951 1157
CNBr	25	1 54 111 137 176 178 268 277 318 567 600 650 668 766 793 825 881 908 945 947 952 971 987 1002 1007
Chymotrypsin-high specificity (C-term to [FYW], not before P)	81	44 48 55 69 74 78 88 112 113 117 125 150 164 172 180 200 254 270 275 285 287 299 316 345 359 376 381 404 420 436 471 477 481 516 549 58 610 626 667 712 723 727 731 764 795 801 813 817 827 856 869 891 898 900 905 910 915 944 961 968 978 997 998 1016 1023 1024 1065 1069 1 1092 1110 1125 1138 1151 1157 1172 1176 1187 1197 1207
Chymotrypsin-low specificity (C-term to [FYWML], not before P)	237	1 11 12 14 15 18 25 38 41 44 47 48 49 51 54 55 62 65 69 74 76 78 79 88 90 93 101 104 112 113 117 119 122 125 132 137 140 144 145 150 156 16 172 173 176 178 180 183 184 200 210 233 249 254 267 268 269 270 275 277 285 287 299 304 316 318 345 349 358 359 369 370 376 381 383 38 395 404 405 406 417 418 420 423 433 436 438 443 448 450 453 469 471 477 480 481 507 509 516 541 542 549 567 584 585 590 600 606 608 61 618 619 626 633 650 654 655 656 657 658 662 666 667 668 672 679 682 683 688 692 703 704 707 712 718 723 727 730 731 747 760 764 766 77 778 782 788 792 795 798 799 801 805 813 814 815 817 825 827 828 833 835 838 844 850 856 858 861 862 869 870 881 883 887 888 891 893 89 905 907 908 910 915 927 944 945 947 952 961 964 968 971 978 979 987 988 997 998 1001 1002 1007 1016 1017 1023 1024 1034 1035 1038 105 1065 1066 1069 1077 1086 1092 1110 1111 1115 1124 1125 1129 1138 1139 1151 1157 1163 1167 1172 1176 1187 1197 1198 1207
Clostripain	60	2 23 53 72 98 108 138 149 165 222 224 244 252 255 297 309 324 334 377 414 427 429 451 494 521 527 531 533 547 574 669 670 671 675 677 68 681 686 705 748 776 803 831 832 836 841 889 932 958 962 973 977 986 999 1031 1052 1068 1100 1121 1199
Formic acid	61	46 75 126 166 171 179 191 230 247 256 262 278 303 314 321 347 368 379 388 393 416 458 460 522 537 577 587 612 761 770 800 807 830 837 8 896 916 942 954 956 974 984 994 1003 1006 1008 1009 1012 1014 1051 1063 1072 1080 1083 1084 1122 1127 1152 1168 1171 1175
Glutamyl endopeptidase	77	26 27 45 59 66 84 97 102 114 134 142 160 204 205 245 257 282 317 319 320 330 344 391 400 412 421 424 455 496 513 519 534 543 545 548 55 561 602 634 685 687 690 697 709 711 734 736 746 749 758 762 804 829 865 866 868 872 884 906 922 928 931 963 967 985 1004 1005 1015 106 1079 1091 1137 1180 1193 1196 1206
Hydroxylamine	5	196 338 639 1053 1184

PeptideCutter - output



Метод скользящего окна

Анализируется последовательность в несколько аминокислот, параметр усредняется по окну. Значение приписывается средней аминокислоте. Output – график

```
Seq. LQAPVLPSDLLSWSCVGAVGILALVSFTCV
     <---*---> Window 1
      <---*---> Window 2
       <---*---> Window 3
```

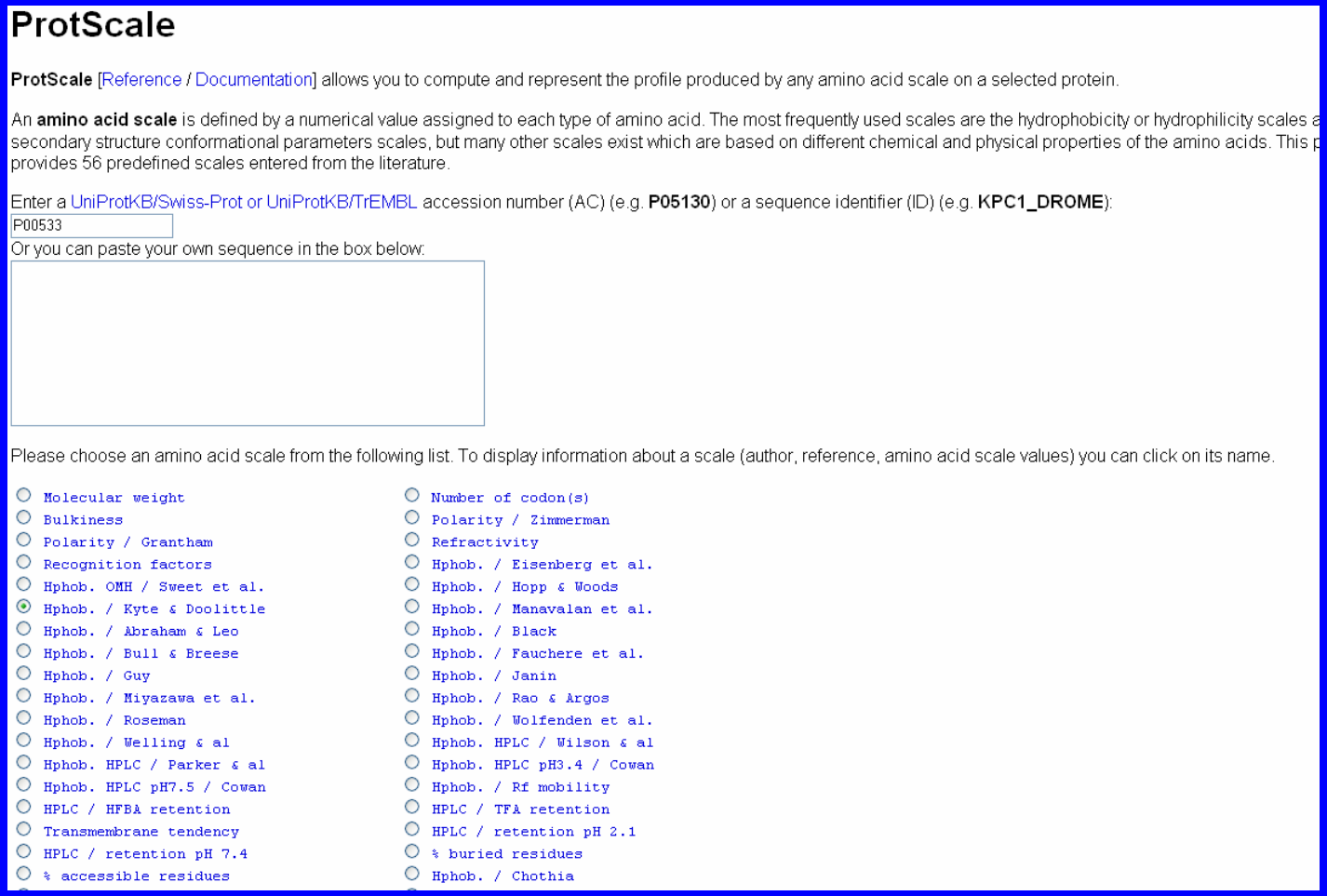
Размер окна должен соответствовать характерному размеру анализируемого свойства (для ТМ – 19!)

Методы, основанные на технике скользящего окна, как правило, не интерпретируют результаты. При интерпретации важно:

✓ Учитывать только очень четко выраженные сигналы

✓ Не зависящие от параметров программы – размера окна, конкретного метода и т.п.

Предсказание трансмембранных сегментов: ProtScale



ProtScale

ProtScale [Reference / Documentation] allows you to compute and represent the profile produced by any amino acid scale on a selected protein.

An **amino acid scale** is defined by a numerical value assigned to each type of amino acid. The most frequently used scales are the hydrophobicity or hydrophilicity scales a secondary structure conformational parameters scales, but many other scales exist which are based on different chemical and physical properties of the amino acids. This p provides 56 predefined scales entered from the literature.

Enter a UniProtKB/Swiss-Prot or UniProtKB/TrEMBL accession number (AC) (e.g. **P05130**) or a sequence identifier (ID) (e.g. **KPC1_DROME**):

P00533

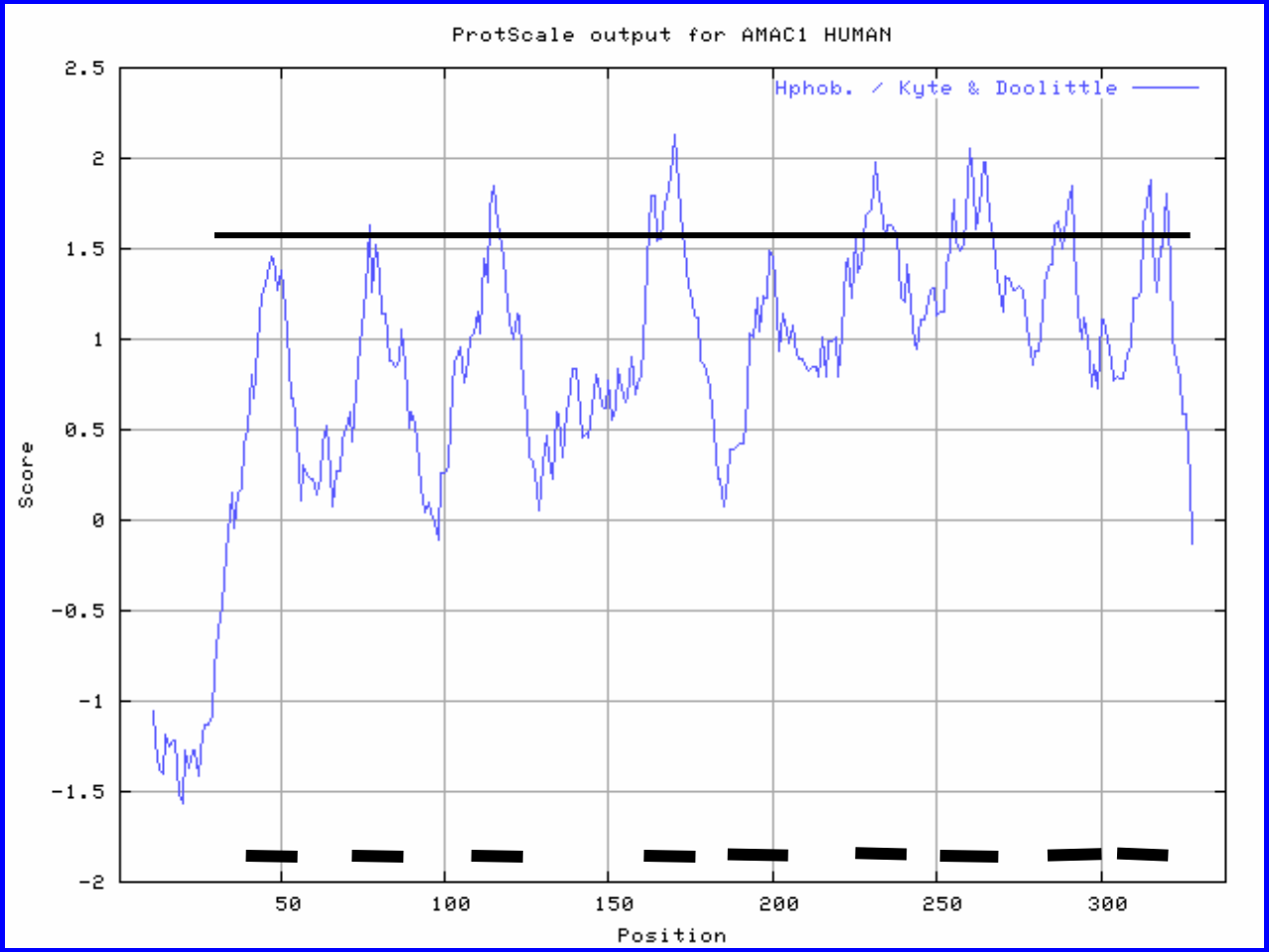
Or you can paste your own sequence in the box below:

Please choose an amino acid scale from the following list. To display information about a scale (author, reference, amino acid scale values) you can click on its name.

- Molecular weight
- Bulkiness
- Polarity / Grantham
- Recognition factors
- Hphob. OMH / Sweet et al.
- (•) Hphob. / Kyte & Doolittle
- Hphob. / Abraham & Leo
- Hphob. / Bull & Breese
- Hphob. / Guy
- Hphob. / Miyazawa et al.
- Hphob. / Roseman
- Hphob. / Welling & al
- Hphob. HPLC / Parker & al
- Hphob. HPLC pH7.5 / Cowan
- HPLC / HFBA retention
- Transmembrane tendency
- HPLC / retention pH 7.4
- % accessible residues
- Number of codon(s)
- Polarity / Zimmerman
- Refractivity
- Hphob. / Eisenberg et al.
- Hphob. / Hopp & Woods
- Hphob. / Manavalan et al.
- Hphob. / Black
- Hphob. / Fauchere et al.
- Hphob. / Janin
- Hphob. / Rao & Argos
- Hphob. / Wolfenden et al.
- Hphob. HPLC / Wilson & al
- Hphob. HPLC pH3.4 / Cowan
- Hphob. / Rf mobility
- HPLC / TFA retention
- HPLC / retention pH 2.1
- % buried residues
- Hphob. / Chothia

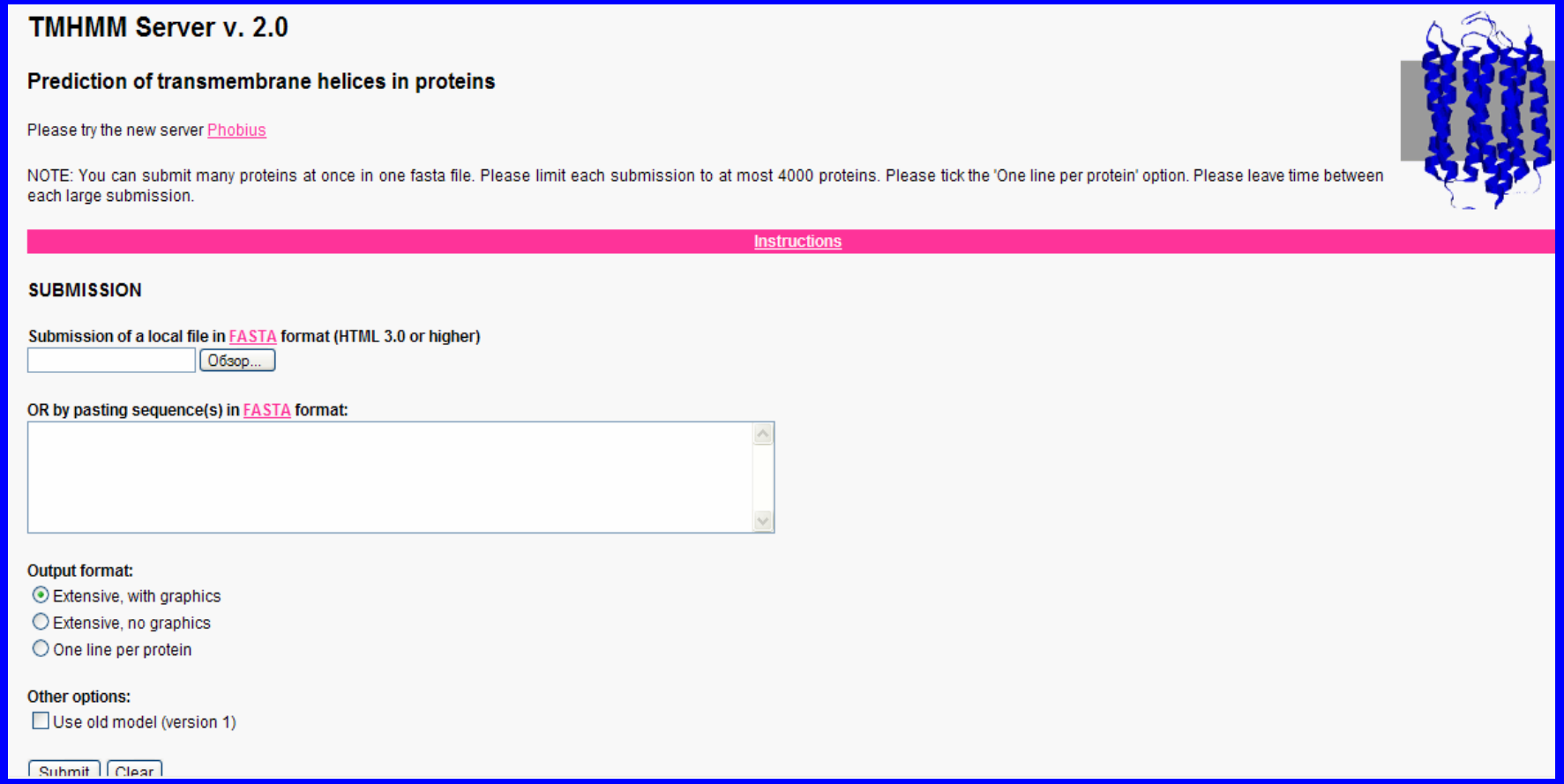
56 аминокислотных шкал (с литературными ссылками), скользящее окно -> выбор ширины окна

ProtScale - output



Правильный порог для метода – 1.6. Здесь находит не все

Более сложное предсказание трансмембранных сегментов: TMHMM



Transmembrane beta barrel prediction: PROFtmb (http://rostlab.org/services/proftmb);
PRED-TMBB (http://biophysics.biol.uoa.gr/PRED-TMBB/);
TBBPred (http://www.imtech.res.in/raghava/tbbpred)

TMHMM - результаты

TMHMM result

HELP with output formats

```
# Q8N808_AMAC1_HUMAN Length: 338
# Q8N808_AMAC1_HUMAN Number of predicted TMHs:  7
# Q8N808_AMAC1_HUMAN Exp number of AAs in TMHs: 174.87557
# Q8N808_AMAC1_HUMAN Exp number, first 60 AAs:  19.01231
# Q8N808_AMAC1_HUMAN Total prob of N-in:        0.45554
# Q8N808_AMAC1_HUMAN POSSIBLE N-term signal sequence
Q8N808_AMAC1_HUMAN      TMHMM1.0        outside      1    35
Q8N808_AMAC1_HUMAN      TMHMM1.0        TMhelix     36    58
Q8N808_AMAC1_HUMAN      TMHMM1.0        inside      59   160
Q8N808_AMAC1_HUMAN      TMHMM1.0        TMhelix    161   179
Q8N808_AMAC1_HUMAN      TMHMM1.0        outside    180   187
Q8N808_AMAC1_HUMAN      TMHMM1.0        TMhelix    188   210
Q8N808_AMAC1_HUMAN      TMHMM1.0        inside     211   222
Q8N808_AMAC1_HUMAN      TMHMM1.0        TMhelix    223   241
Q8N808_AMAC1_HUMAN      TMHMM1.0        outside    242   249
Q8N808_AMAC1_HUMAN      TMHMM1.0        TMhelix    250   272
Q8N808_AMAC1_HUMAN      TMHMM1.0        inside     273   278
Q8N808_AMAC1_HUMAN      TMHMM1.0        TMhelix    279   301
Q8N808_AMAC1_HUMAN      TMHMM1.0        outside    302   309
Q8N808_AMAC1_HUMAN      TMHMM1.0        TMhelix    310   328
Q8N808_AMAC1_HUMAN      TMHMM1.0        inside     329   338
```

plot in postscript, script for making the plot in gnuplot, data for plot

---

Находит только 7! TMs

```
FT   CHAIN         1    338
FT   TRANSMEM     37     57
FT   TRANSMEM     67     87
FT   TRANSMEM    105    125
FT   TRANSMEM    160    180
FT   TRANSMEM    185    205
FT   TRANSMEM    221    241
FT   TRANSMEM    250    270
FT   TRANSMEM    281    301
FT   TRANSMEM    305    325
```

TMHMM предсказывает сегменты, а также топологию межсегментных участков

Домены

- Домен – независимая глобулярная единица в белке. Более функционально – часть белка, обладающая активностью (если отрезать, например). Как правило, каждый домен играет свою роль в функции белка (связывает ион или ДНК, содержит активный сайт и т.п.)
- Только небольшая часть известных доменов была изучена экспериментально, остальные описаны как сходные части гомологичных белков
- Очень сложно четко определить домен и его границы => существует много подходов и различных доменных коллекций. Какую выбрать?

История коллекций доменов

- ✓ 1980ые – PROSITE: ручная выборка паттернов в белках, определяющих функцию
- ✓ 1987 – доменный профайл (Gribskov):
  position specific scoring schema – это вероятность для каждой аминокислоты находиться в данной позиции домена
- ✓ начало 1990х – BLOCKs, PRINTs, Prodom…
- ✓ PfamA – коллекция профайлов, курированная вручную (сейчас также использует HMM)

3 сервера для поиска доменов

- ✓InterProScan

http://*www.ebi.ac.uk/**InterPro**Scan*

- ✓CD (Conserved Domain) server (NCBI)

http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

- ✓Pfscan (Motif Scan)

http://hits.isb-sib.ch/cgi-bin/PFSCAN

InterPro

InterPro is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences.

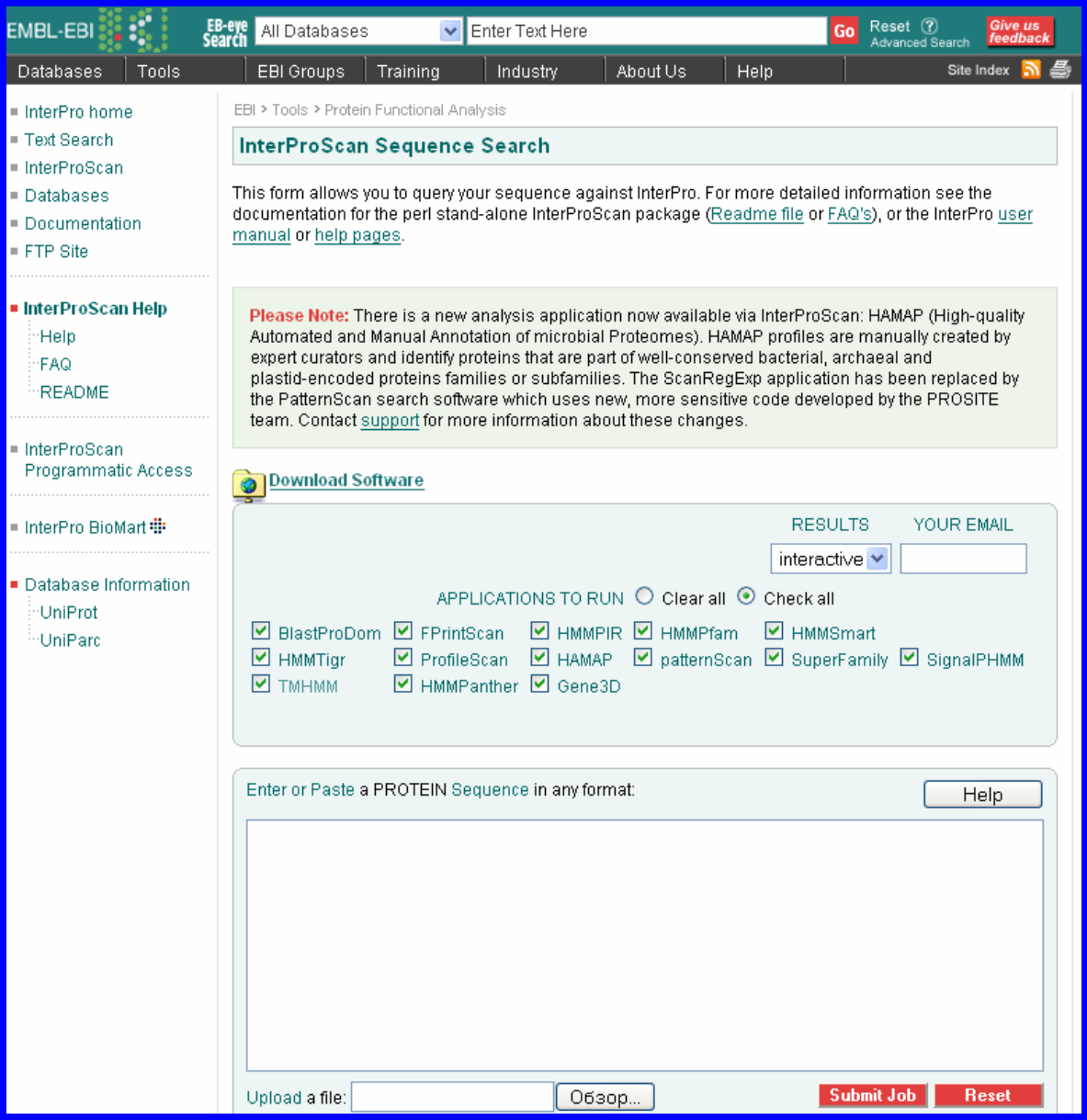
Классификация базируется на первичных классификациях целого ряда баз данных функциональных доменов и семейств, объединяет всю доступную информацию

С 2001 года – Release 23.0: 76.4% UniProt

Как это происходит

Каждое InterPro семейство объединяет первичные семейства других баз данных, описывающие один и тот же домен; включает все белки, принадлежащие хотя бы одной из первичных баз. Документация IP семейства подробно описывает функцию и структуру соответствующей белковой подписи.

Поиск доменов: InterProScan



InterProScan - результаты

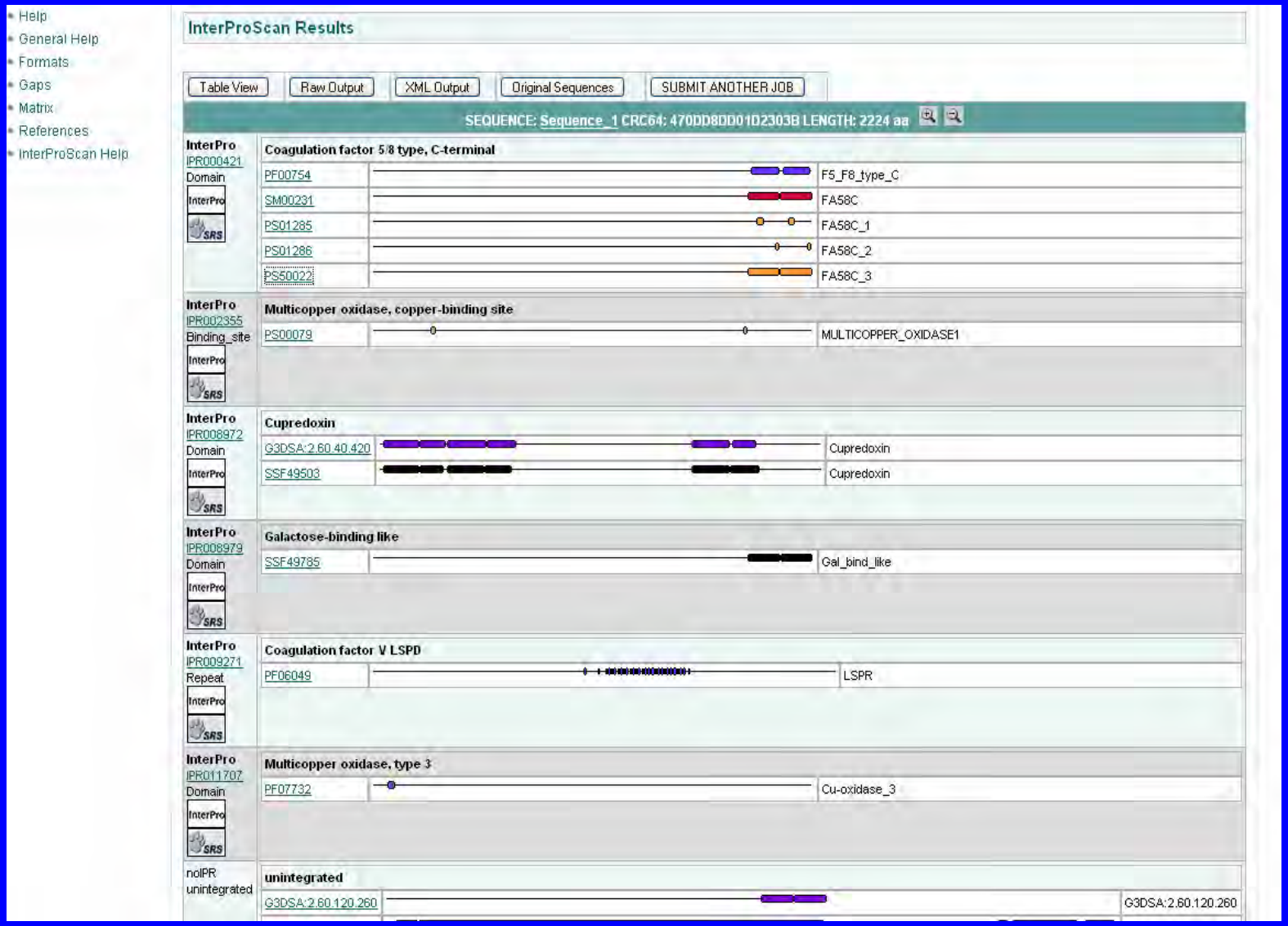


Table View

- Help
- General Help
- Formats
- Gaps
- Matrix
- References
- InterProScan Help

InterProScan Results

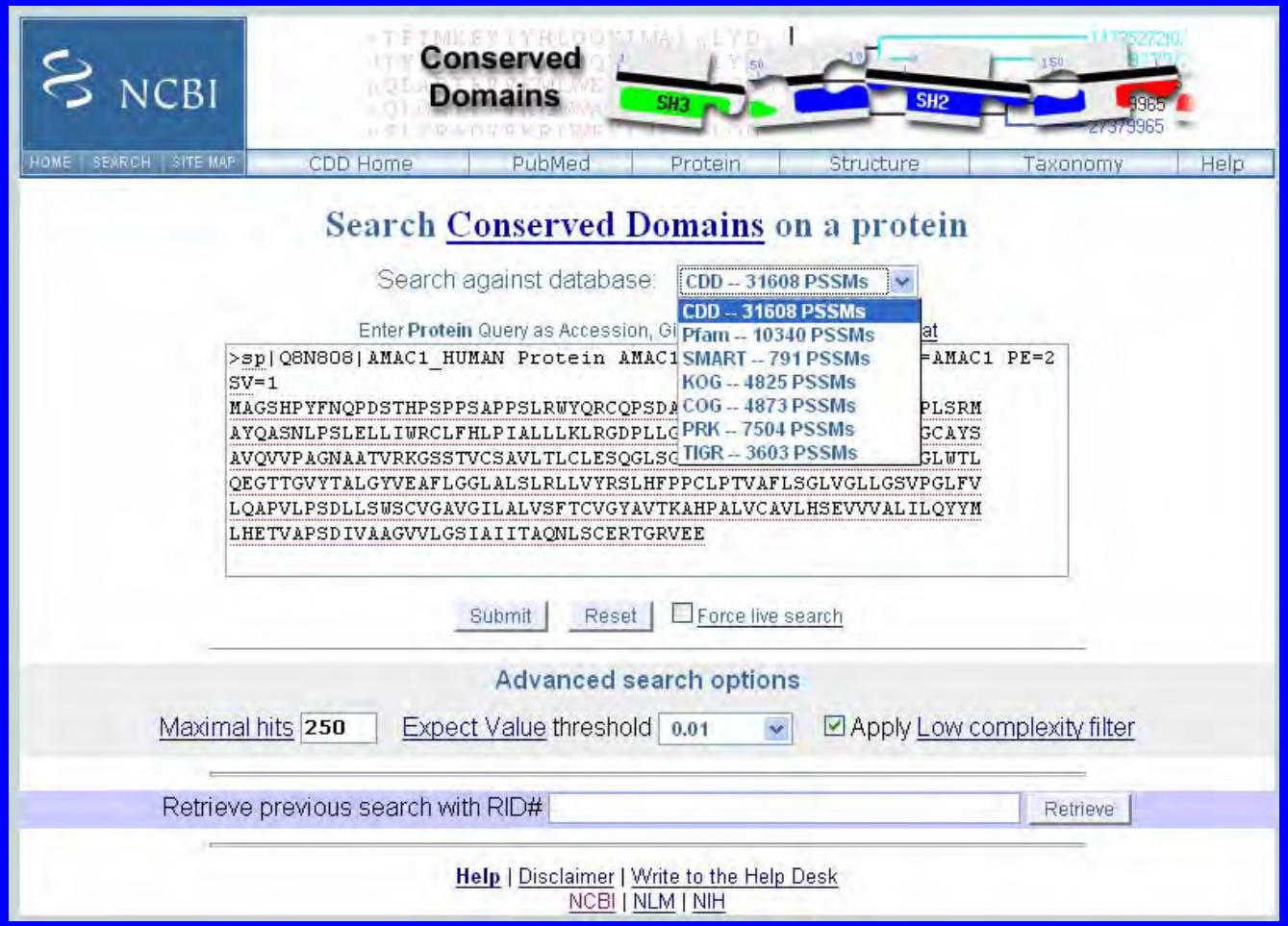
Picture View | Raw Output | XML Output | Original Sequences | SUBMIT ANOTHER JOB

SEQUENCE: Sequence_1 CRC64: 470DD8DD01D2303B LENGTH: 2224 aa

InterPro IPR000421 Domain	**Coagulation factor 5/8 type, C-terminal**			
	PFAM	PF00754	*F5_F8_type_C*	1.40000506907309E-33 [1922-2058]T 5.30001985225048E-41 [2081-2218]T
	SMART	SM00231	*FA58C*	5.7999999999995E-42 [1906-2061]T 2.49999999999978E-30 [2065-2221]T
	PROFILE	PS01285	*FA58C_1*	0.0 [1947-1980]T 0.0 [2111-2140]T
	PROFILE	PS01286	*FA58C_2*	0.0 [2045-2061]T 0.0 [2205-2221]T
	PROFILE	PS50022	*FA58C_3*	0.0 [1907-2061]T 0.0 [2066-2221]T
Parent	IPR008979			
Children	no children			
Found in	IPR012111 IPR014648 IPR014707			
Contains	no entries			
GO terms	Biological Process: cell adhesion (GO:0007155)			
InterPro IPR002355 Binding_site	**Multicopper oxidase, copper-binding site**			
	PROFILE	PS00079	*MULTICOPPER_OXIDASE1*	0.0 [304-324]T 0.0 [1880-1900]T
Parent	no parent			
Children	no children			
Found in	IPR006376 IPR008972 IPR010532 IPR011706 IPR011707 IPR014707			
Contains	no entries			
GO terms	Molecular Function: copper ion binding (GO:0005507)			
InterPro IPR008972 Domain	**Cupredoxin**			
	GENE3D	G3DSA:2.60.40.420	*Cupredoxin*	5.69999708784002E-63 [29-209]T 3.99998544139379E-43 [210-334]T 2.39999798157241E-61 [347-542]T 5.500017542664309E-40 [543-689]T 1.1000015067164299E-63 [1577-1767]T 2.0000007521745 6E-5 [1781-1900]T
	SUPERFAMILY	SSF49503	*Cupredoxin*	2.6E-40 [29-208]T 1.1E-31 [209-324]T 2.5000000000000004E-36 [348-532]T 1.1000000000000001E-35 [539-665]T 4.5E-42 [1574-1763]T 7.4999999999999996E-34

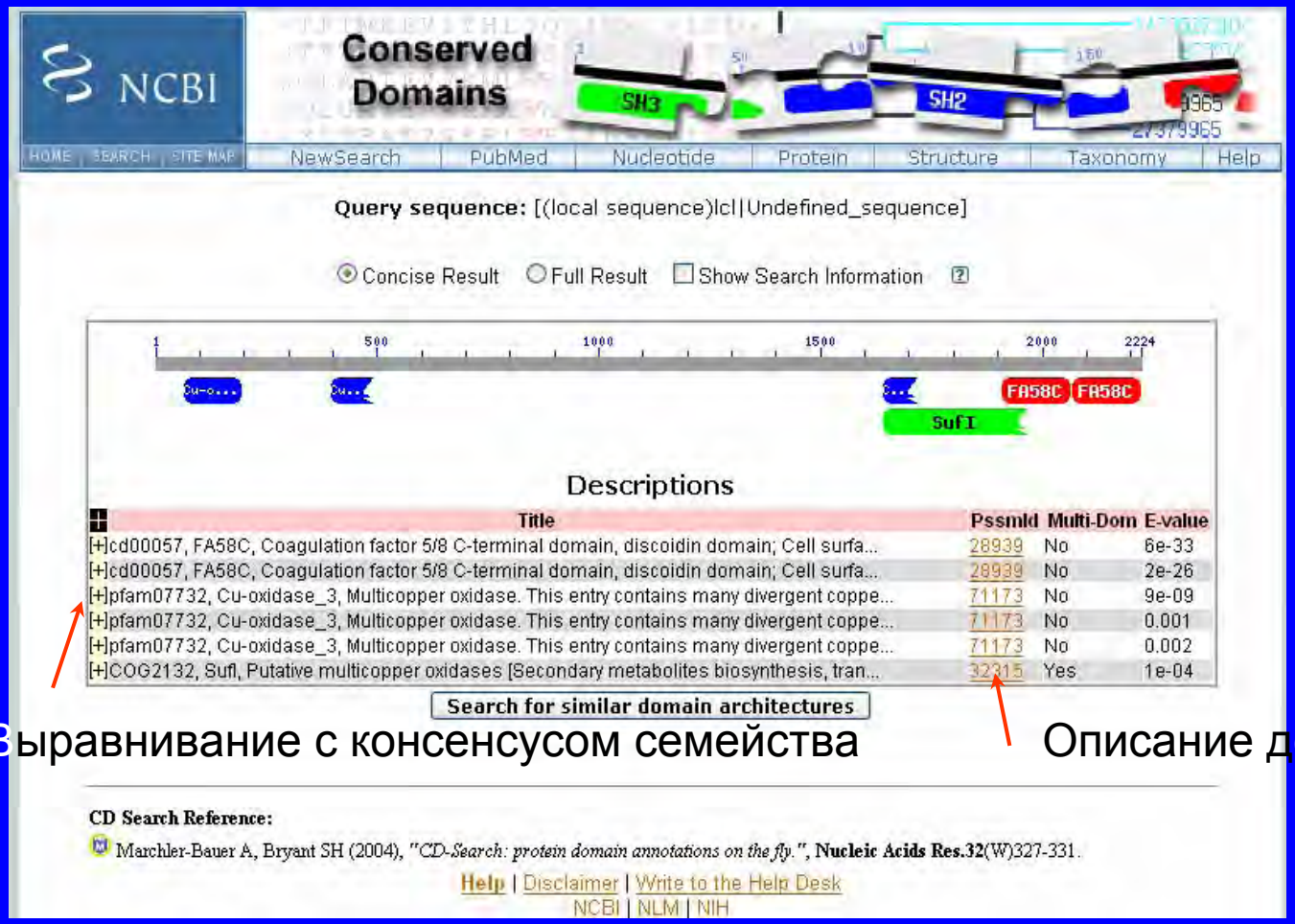
Готово

CD server



Input - Accession number, gi или последовательность в FASTA формате

CD server – output



Выравнивание с консенсусом семейства

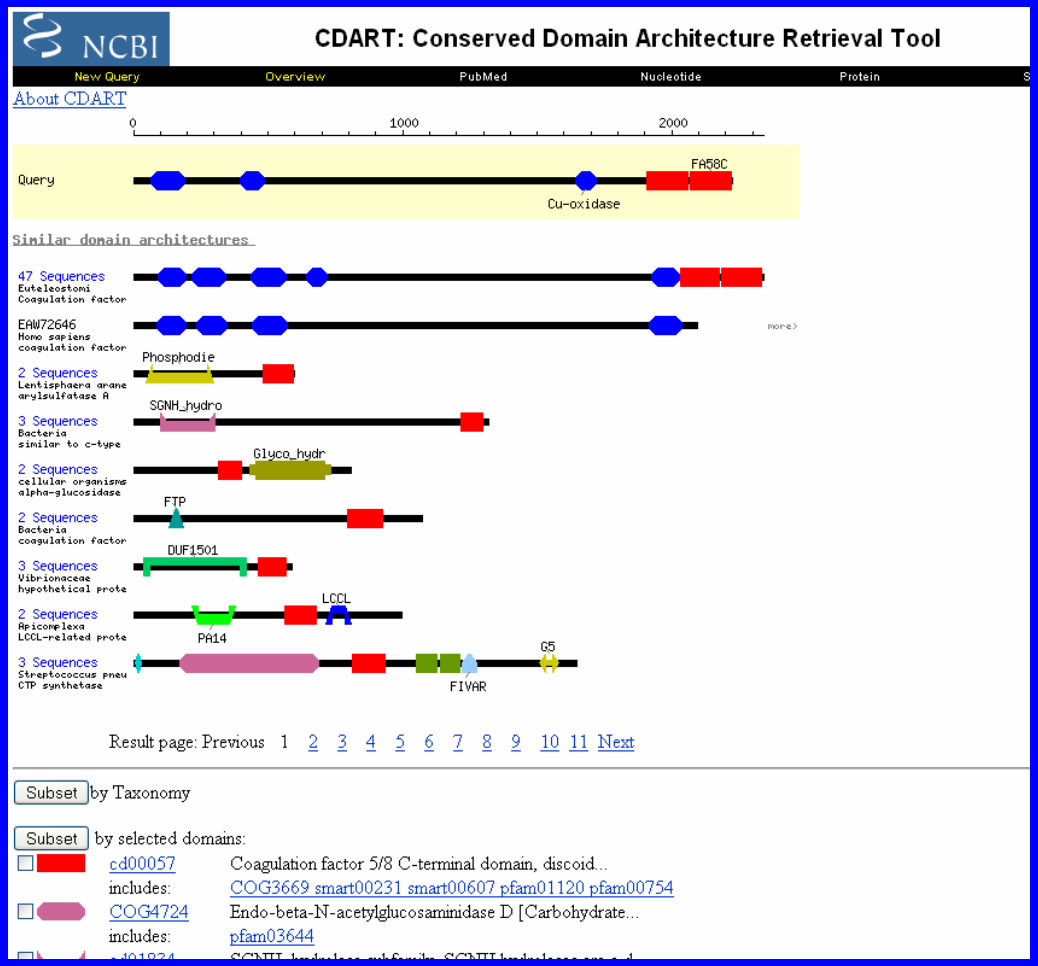
Описание домена

Красный – SMART, синий – Pfam, зеленый – COGs

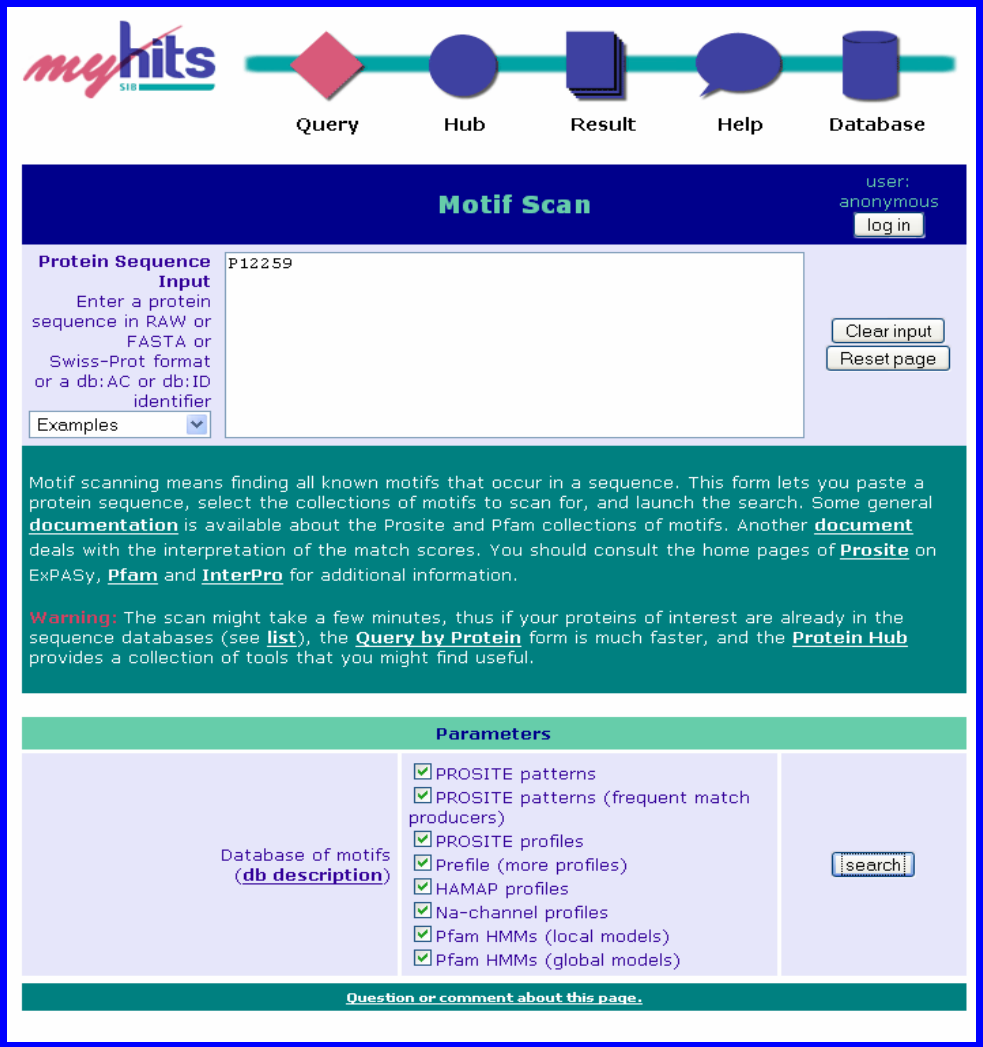
Рваные концы указывают на неполные домены!!!!

Курсор в графической части – краткое описание функции домена

CDART – поиск белков с аналогичной доменной структурой

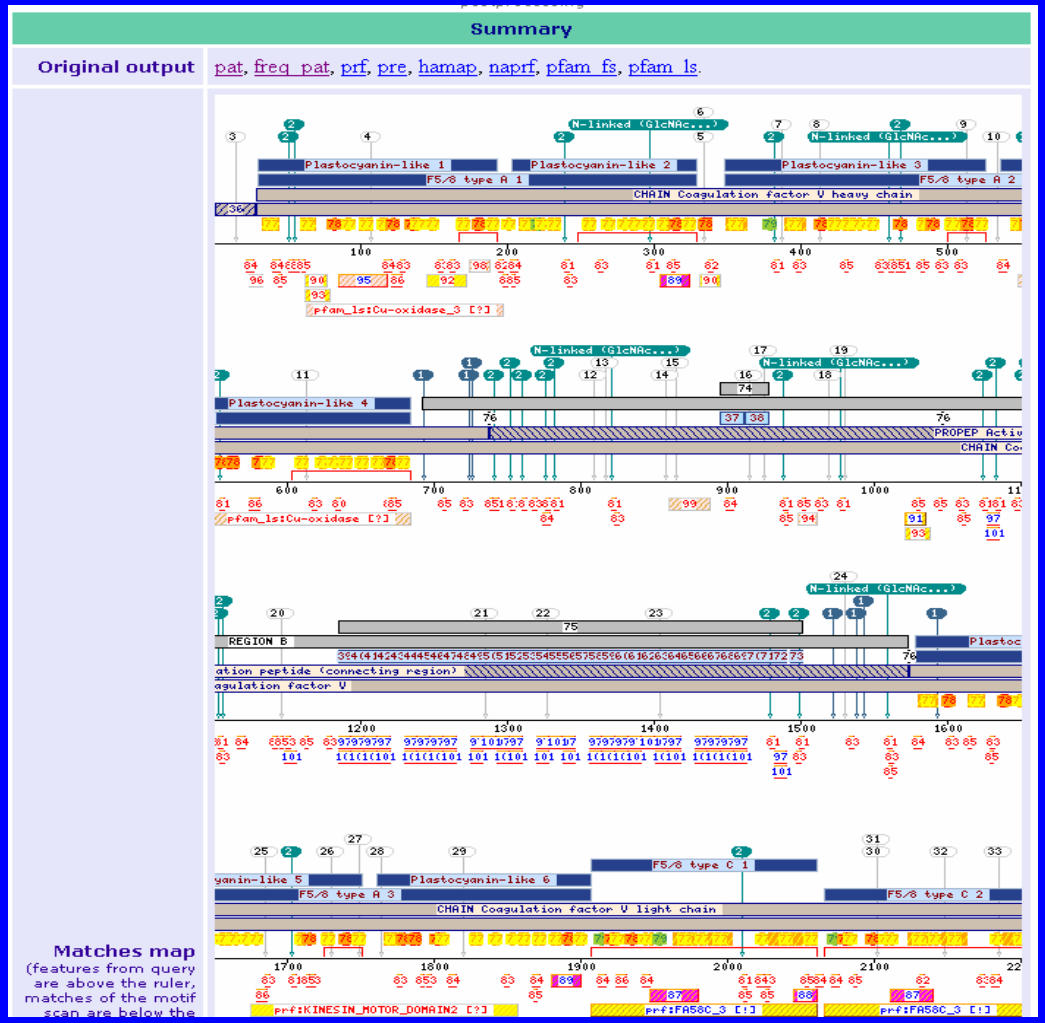


Pfscan



Как правило, работает несколько минут

Pfscan - output



Особенности вывода Pfscan

- Схема – легенда, как всегда под рисунком
- За легендой следует таблица с локализацией доменов
- Далее расшифровка каждого хита – с оценкой вероятности: ? или !
- Затем следует графическая схема для каждого хита и scores (высокий score = хороший хит)

“Match detail” (или графическая схема)

Alignment of a Sequence on a Profile

The pairwise alignment below corresponds to the one obtained when the PDI_ASPNG sequence is searched with the THIOREDOXIN_2 profile. For the sake of the textual representation, the profile positions were symbolized by the residues of the "consensus" sequence of the multiple sequence alignement from which the profile was derived. This alignment is not fundamentally different from the one considered before

```
consensus    1 XVXVLSDENFDEXVXDSDKPVLVDFYAPWCGHCRALAPVFEELAEEYK----DBVKFVKV  -48
               : :          : : :  ::  : : ::::: : :    : :  :         :
PDI_ASPNG  360 PVTVVVAHSYKDLVIDNDKDVLLEFYAPWCGHCKALAPKYDELAALYAdhpdLAAKVTIA  -97

consensus   57 DVDENXELAEEYGVRGFPTIMFF--KBGEXVERYSGARBKEDLXEFIEK              -1
                 :            : ::                : : :  : :  :
PDI_ASPNG  420 KID-ATANDVPDPITGFPTLRLYpaGAKDSPIEYSGSRTVEDLANFVKE              -49
```

but the textual representation does not reveal the additional information carried on by the profile scoring system, that eventually makes the identification by the profile so "informative". The alternative graphical representation of this alignment reveals much of this extra information.

- In strong contrast to the previous example, the scoring system is heavily position-dependent: The area of every red rectangle corresponds to the score attributed by the profile for the