Hidden Markov Models (HMMs) and Profiles

Swiss Institute of Bioinformatics (SIB)

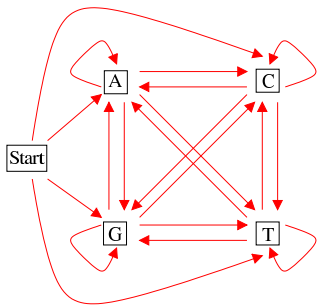
26-30 November 2001

Markov Chain Models

A Markov Chain Model is a succession of **states** S_i $(i = 0, 1, ...)$ connected by **transitions**. Transitions from state S_i to state S_j has a probability of P_{ij}.

An example of Markov Chain Model:

- Transition probabilities:
  - $P(A|G) = 0.18$, $P(C|G) = 0.38$, $P(G|G) = 0.32$, $P(T|G) = 0.12$
  - $P(A|C) = 0.15$, $P(C|C) = 0.35$, $P(G|C) = 0.34$, $P(T|C) = 0.15$



Markov Chain Models

Given a sequence x of length L, we can ask how probable the sequence is given a Markov Chain Model:

$$P(x) = (P(x_L), P(x_{L-1}), ..., P(x_1)) =$$
$$P(x_L|x_{L-1}, ..., x_1)P(x_{L-1}|x_{L-2}, ..., x_1)...P(x_1)$$

Key property of a Markov Chain (of order 1):

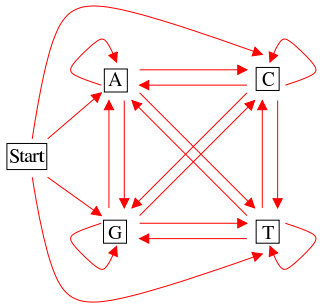
$$P(x_i|x_{i-1}, x_{i-2}, ..., x_1) = P(x_i|x_{i-1}).$$

Therefore:

$$P(x) = P(x_L|x_{L-1})P(x_{L-1}|x_{L-2})...P(x_1) =$$
$$P(x_1)\prod_{i=2}^{L} P(x_i|x_{i-1})$$

What all this stuff means?

Given a Markov Chain Model M where all transition probabilities are known:



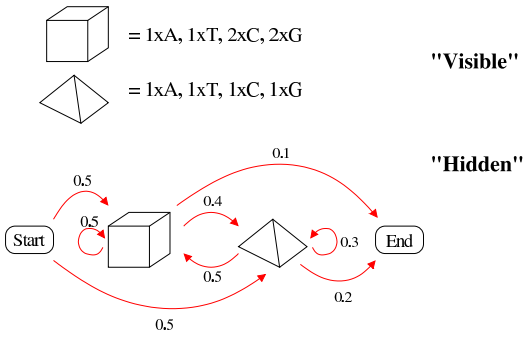
The probability of sequence $x = GCCT$ is:

$$P(GCCT) = P(T|C)P(C|C)P(C|G)P(G)$$

Hidden Markov Models

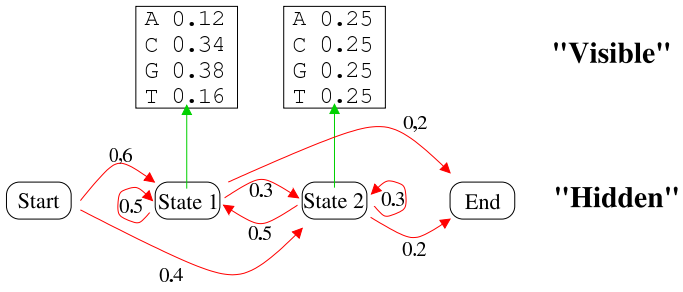
Hidden Markov Models (HMMs) are like Markov Chain Models: a finite number of **states** connected between them by **transitions**.

But the major difference between the two is that the states of the Hidden Markov Models are not a symbol but a **distribution** of symbols. Each state can **emit** a symbol with a probability given by the distribution.



Hidden Markov Model

Example of a simple Hidden Markov Model:



START 1 1 1 1 2 2 1 1 1 2 END

G C A G C T G G C T

Hidden Markov Models

The parameters of the HMMs:

- **Emission probabilities.** This is the probability of emitting a symbol x from an alphabet α being in state q.

$$E(x|q)$$

- **Transition probabilities**. This is the probability of a transition from to a state r being in state q.

$$T(r|q)$$

Parameter estimation

How we can determine the parameters (emission and transition probabilities) of our model?

To estimate the parameters we use Maximum Likelihood Estimation.

Estimation of the probability of observing one symbol from an alphabet α (in our case $\alpha = \{A, C, G, T\}$):

- Given a set of sequences:

gccgcgcttg
gcttggtggc
tggccgttgc

- the maximum likelihood estimates are:

$$P(A) = \frac{0}{30} = 0 \qquad P(G) = \frac{13}{30} = 0.433$$

$$P(C) = \frac{9}{30} = 0.3 \qquad P(T) = \frac{8}{30} = 0.267$$

Parameters estimation

To avoid to set probabilities to 0 we can use different methods:

- For example add pseudo-counts:

$$P(A) = \frac{0+1}{34} = 0,029 \quad P(G) = \frac{13+1}{34} = 0.412$$

$$P(C) = \frac{9+1}{34} = 0.294 \quad P(T) = \frac{8+1}{34} = 0.265$$

A general form for priors estimation:

$$P(A) = \frac{n_a + p_A m}{(\sum_i n_i) + m}$$

where m is the number of virtual instances (pseudo-counts) and p_A is the prior probability of A.

Parameters estimation

Transition probabilities estimation between the symbols of the alphabet α (for example $\alpha = \{A, C, G, T\}$):

- Transition probabilities are estimated for each observed couple of symbols:

$$p_{ij} = \frac{n_{ij}}{\sum_{j'} n_{ij'}}$$

- where p_{ij} is the probability of transition from symbol i to symbol j, and n_{ij} is the number of transition from symbol i to symbol j observed.
- Is possible to add pseudo-counts as described previously.

For example for transition A $\longrightarrow$ G:

- Count all transitions A $\longrightarrow$ A, A $\longrightarrow$ C, A $\longrightarrow$ G, A $\longrightarrow$ T in CpG sequences and non-CpG sequences.
- Transition probability is then calculated:

$$p_{AG} = \frac{n_{AG}}{n_{AA} + n_{AC} + n_{AG} + n_{AT}}$$

An example

Distinguish CpG islands from other sequences regions.

- Two models are required:
  - a model to represent CpG islands
  - a **null model** to represent the other regions

A sequences x is scored for being a CpG island:

$$score(x) = log\frac{P(x|CpGmodel)}{P(x|nullmodel)}$$

Parameters for the CpG island and null models:

CpG	A	C	G	T	Null	A	C	G	T
A	0.18	0.27	0.43	0.12	A	0.18	0.27	0.43	0.12
C	0.17	0.37	0.27	0.19	C	0.17	0.37	0.27	0.19
G	0.16	0.34	0.38	0.12	G	0.16	0.34	0.38	0.12
T	0.08	0.36	0.38	0.18	T	0.08	0.36	0.38	0.18

HMMs algorithms

Three important questions can be answered by three algorithms.

How likely is a given sequences under a given model? This is the scoring problem and it can be solved using the **Forward algorithm**.

What is the most probable path between states of a model given a sequence? This is the alignment problem and it is solved by the **Viterbi algorithm**.

How can we learn the HMM parameters given a set of sequences? This is the training problem and is solved using the **Forward-backward algorithm** and the **Baum-Welch expectation maximization**.

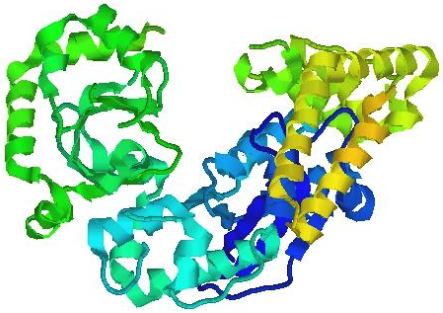
For details about these algorithms see:
Durbin, Eddy, Mitchison, Krog.
Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.
Cambridge University Press, 1998.

HMMs and protein domains

HMMs are very appropriate for modelization of protein domains.

What are protein domains:

- Domains are discrete structural units.
- Defined by structure.
- Domains are functional units.



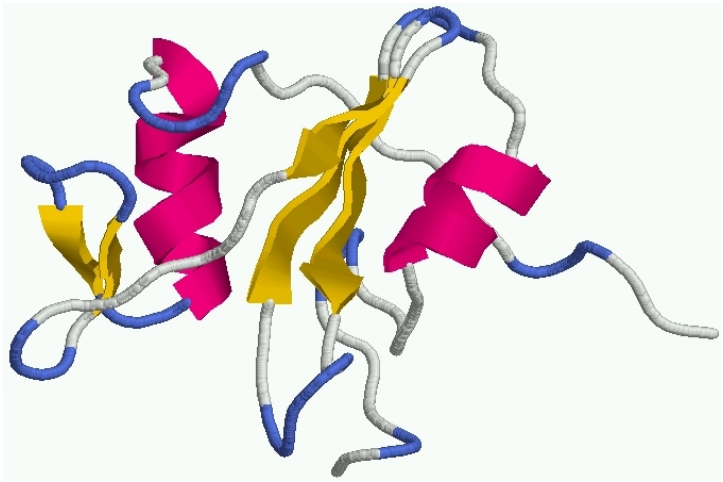
HMMs and protein domains

Multiple alignments are used as training set for the building of HMMs.

sh2.stk

```
ABL1_CAEEL   194 WYFGKISRSDSEAILGS..GITGSFLVRESET.SIGQ...YTLSVRHDG.......RVFHYRINVDNTE......KMFITQEVK..FRTLGELVHH 269
P85A_HUMAN   624 WNWGSSNRNKAENLLRG..KRDGTFLVRESS..KQGC...YACSVVVDG.......EVKHCVINKTATG........YGFAEPYNLYSSLKELVLHY 698
DRK_DROME     60 WYYGRITRADAEKLLSN..KHEGAFLIRIRES.SPGD...FSLSVKCPD.......GVCHTKVLRDAQS........KFFLWVVK..FNSLNELVEYH 134
NCK_HUMAN    282 WYYGKVTRHQAEMALNE.RGHEGDFLIRDSES.SPND...FSVSLKAQG.......KNKHIKVQLKET.........VYCIGQRK..FSTMEELVEHY 356
BLK_MOUSE    117 WFFRTISRKDAERQLLAPMNKAGSFLIRESES.NKGA...FSLSVKDITTQG...EVWKHYKIRSLDNG......GYYISPRIT..FPTLQALVQHY 198
YES_XIPHE    159 WYFGKLSRKDTERLLLLPGNERGTFLIRESET.TKGA...YSLSLRDWDETK..GDNCKHYKIRKLDNG......GYYITTRTQ..FMSLQMLVKHY 241
STK_HYDAT    126 WFFGNVKRAEAEKKLMVRGLPSGTFLIRQSET.AVGN...FSLSVRDGD.......SVKHYRIRKLDTG......GFFITTRAF..FNSLYELVQHY 203
SRK1_SPOLA   122 WFLGKIKRVEAEKMLNQSRNQVGSFLIRDSET.TPGD...FSLSVKDQD.......FVRHYRVRRLEDG......SLFVTRRST..FQTLHELVLHY 199
SRC1_DROME   162 WFFENVLRKEAIKLLLAEENPRGTFLVRFSEH.NPNG...YSLSVKDWEDGR..GYHVKHYRIKPLDNG......GYYIATNQT..FPSLQALVMAY 244
CRKL_HUMAN    14 WYMGPVSRQEAQTKLQG..QRHGMFLVRDSST.CPGD...YVLSVSENS.......RVSHYIINSLPNR.......FFKIGDQE..FDHLPALLEFY 88
CSK_CHICK     82 WFHGKITREQAERLLYP..PETGLFLVRESTN.YPGD...YTLCVSCEG.......KVEHYRIIYSSSK.......LSIDEEVY..FENLMQLVEHY 156
MATK_HUMAN   122 WFHGKISGQEAVQQLQP..PEDGLFLVRESAR.HPGD...YVLCVSFGR.......EVIHYRVLHRDG........HLTIDEAVF..FCNLMDMVEHY 196
CSW_DROME      6 WFHPTISGIEAEKLLQE.QGFDGSFLARLASS.NPGA...FTLSVRRGN.......EVTHIKIQNNGDF.......FDLYGGEK..FATLPELVQYY 81
CSW_DROME    111 WFHGNLSGKEAEKLLLE.RGKNGSFLVRESQS.KPGD...FVLSVRTDD.......KVTHVMIRWQDKK.......WDVGGGES..FGTLSELLDHY 186
GTPA_HUMAN   181 WYHGKLDRTIAEERLRQ.AGKSGSYLLRESDR.RPGS...YVLSFLSQMN.......VYNHFRIIAMCGD.......KYIGG.RR..FSSLSDLLSYY 256
PIP4_RAT     668 WYHASITRAQAEHMLMR.VPRDGAFLVRKRNE..PNS...YALSFRAEG.......KIKHCRVQQEGQT.......VMLGNSE...FDSLVDLISYY 741
KSYK_PIG     163 WFHGKISRDESHQIVLIGSKTNGKFLIRARDN...GS...YALSLHEGG.......KVLHYRIDKDKTG.....KLSIPGGKN..FDTLWQLVEHY 238
ZA70_HUMAN   163 WYHSSLTREEAERKLYSGAQTDGKFLLRPRKE..QGT...YALSLIYGK.......TVYHYLISQDKAG.....KYCIPEGTK..FDTLWQLVEYL 239
VAV_MOUSE    671 WYAGPMERAGAEGLLTN..RSDGTYLVRQRVK.DTAE...FALSIKYNV.......EVKHIKIMTSEG.......LYRITEKKA.FRGLLELVEFY 745
KSYK_HUMAN    15 FFFGNITREEAEDYLVQGGMSDGLYLLRQSRN.YLGG...FALSVAHGR.......KAHHYTIERELNG.....TYAIAGGRT..HASPADLCHYH 92
SPK1_DUGTI   100 EAWREIQRWEAEKSLMKIGLQKGTYLIRPGR..KENS...YALSVRDFDEKKK.ICIWKHFQLKTLQDEK.....GISYSVNIRN.FPNILTLIQFY 184
BTK_HUMAN    281 WYSKHMTRSQAEQLLKQ.EGKEGGFIVRDSS..KAGK...YIVSVFAKSTGDP.QGVIRHVVVCSTPQS.....QYYLAEKHL..FSTIPELINNYH 362
TEC_MOUSE    246 WYCRNTNRSKAEQLLRT.EDKEGGFMVRDSS..QPGL...YIVSLYTKFGGEG.SSGFRHVHHKETATSP....KKVYLAEKHA..FSSIPEIINFYH 329
ITK_HUMAN    239 WYMKSISRDKAEKLLLD.TGKEGAFWVRDSR..TAGT...YIVSVFTKAVVSENNPCLKHHHIKETNDNP....KRYYVAEKYV..FDSIPLLINYH 323
TXK_HUMAN    150 WYHRNITRNQAEHLLRQ..ESKEGAFIVRDSR..HLSS...YTISVFMGARKST.EAALKHYQIKKNDSG.....QWYVAERHA..FQSIPELIWYH 231
SRC2_DROME   214 WYVGYMSRQRAESLLKQ.GDKEGCFVVRKSS..TKGL...YTLSLHTKVPQ....SHVKHYHIKQNARC......EYYLSEKHC..FCTIPDLINYH 292
FER_HUMAN    460 WYHGAIPRIEAQELLK....KQGDFLVRESHG.KPGE...YVLSVYSDG.......QRHHIIQYVDNM........YRFEGTG..FSNIPQLIMH 531
FPS_DROME    438 WFHGVLPREEVVRLLNN....DGDFLVRETIRNEESQ...IVLSVCWNGH........KHTIVQTTGEG.......NFRFEGPP..FASIQELIMMQ 510
GTPA_HUMAN   351 WFHGKISKQEAYNLLMT.VGQVCSFLVRPSDN.TPGD...YSLYFRTNE.......NICRTKICPTPN.......NQFMMGGRY..YNSIGDILHY 426
YKF1_CAEEL    20 YFHGLIQREDVFQLLDN....NDYVVRLSDP.KPGEPRSYILSVMFNNKLDE.NSSVKHFVINSVENK.......KFVNNNMS..FNTIQQMLSHY 101
SPT6_YEAST  1258 YFPFNGR.QAEDYLRS..KERGEEFVIRQSSR.GDDH...LVLSWKLDKD......LFQHIDLQELEKENPLALGKVLIVDNQK..YNDLDQILVEY 1339
```

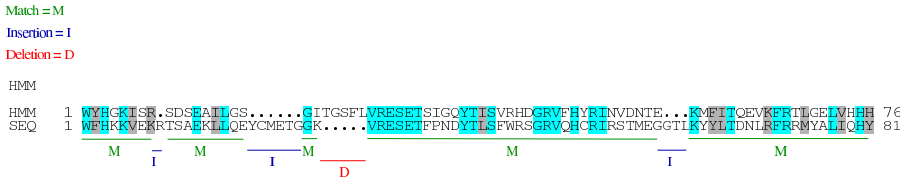
HMMs and protein domains



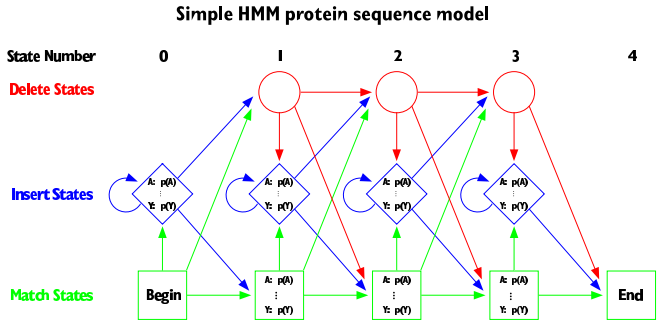
HMMs and protein domains

A model based on three types of states is appropriate to modelize biological sequences:

- **match state**: Emits a symbol corresponding to a match.
- **insert state**: Emits a symbol between matches.
- **delete state**: Non-emitting silent state.



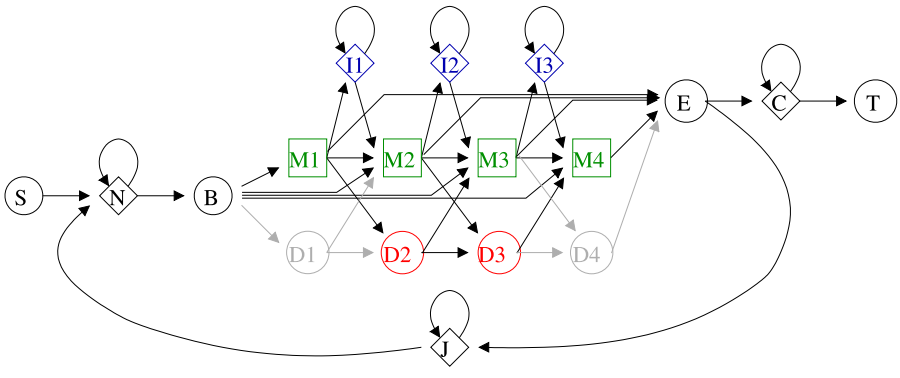
HMMs and protein domains



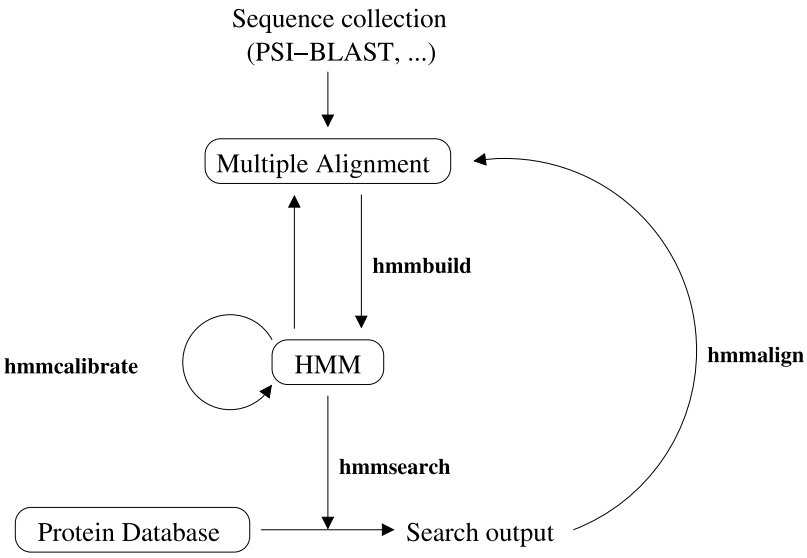
HMMER2

Package developed by Sean Eddy (http://hmmer.wustl.edu/).

HMMER uses the "Plan 7" architecture:



HMMER2



Generalized profiles

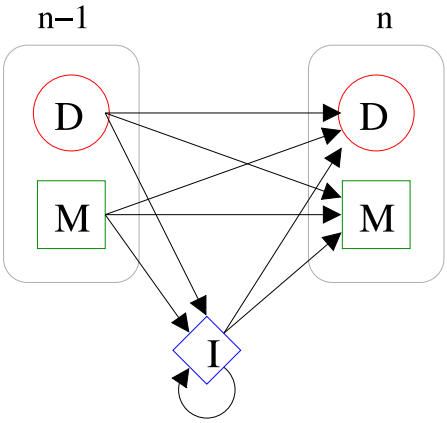
Generalized profiles combine aspects of Position Specific Score Matrices (PSSMs) and Hidden Markov Models (HMMs).

Generalized profiles are composed of alternating match and insert positions. Scores are associated with each position.

- The match position gives a residue specific match extension score and a deletion extension score.
- The insert position gives residue specific insertion scores.
- Scores are also associated to all possible transitions.

There is an equivalence between the structure of generalized profiles and a linear HMMs.

Generalized profiles



Pftools

The package has been developed by Philipp Bucher (http://www.isrec.isb-sib.ch/ftp-server/pftools/).

The package contains:

- pfmake for building a profile starting from multiple alignments.
- pfsearch to search a profile against a protein database.
- pfscan to search a protein against a profile database.

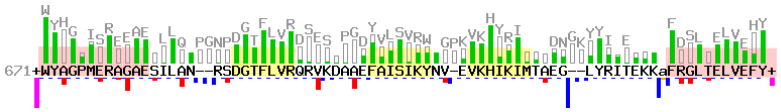
Two tools have been created to translate profiles from and into HMMs:

- htop: HMM to profile.
- ptoh: profile to HMM.

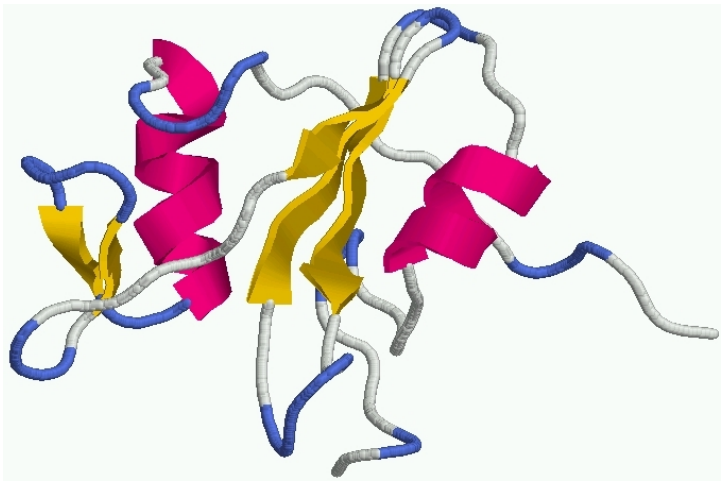
HMM-profiles and Generalized profiles

The meaning of the position specific scores in the generalized profiles:

The emission probabilities of HMM-profiles can be translated into a scoring system as for the generalized profiles:



HMMs and protein domains



Protein domain databases

A non-exhaustive list of protein domain databases:

- Pfam
  - http://www.sanger.ac.uk/Pfam.
  - Collection of protein domains and families (3071 entries in Pfam release 6.6).
  - Uses HMMs (HMMER2).
  - Good links to structure, taxonomy.
- PROSITE
  - http://www.expasy.ch/prosite.
  - Collection of motifs, protein domains, and families (1494 entries in Prosite release 16.51).
  - Uses generalized profiles (Pftools) and patterns.
  - High quality documentation.

Protein domain databases

A non-exhaustive list of protein domain databases (continue):

- Prints
  - http://bioinf.man.ac.uk/dbbrowser/PRINTS.
  - Collection of conserved motifs used to characterize a protein.
  - Uses fingerprints (conserved motif group).
  - Very good to describe sub-families.
  - Release 32.0 of PRINTS contains 1600 entries, encoding 9800 individual motifs.
- ProDom
  - http://prodes.toulouse.inra.fr/prodom/doc/prodom.html.
  - Collection of protein motifs obtained automatically using PSI-BLAST.
  - Very high throughput ... but no annotation.
  - ProDom release 2001.2 contains 101957 families (at least 2 sequences for family).
- ...

InterPro

InterPro is an attempt to group a number of protein domain databases:

- Pfam
- PROSITE
- PRINTS
- ProDom
- SMART
- TIGRFAMs

InterPro try to have and maintain a high quality annotation.

Very good accession to examples.

InterPro web site: http://www.ebi.ac.uk/interpro.

A stand alone package (iprscan) and the database are available for UNIX platforms to run a complete Interpro analysis: ftp://ftp.ebi.ac.uk/pub/databases/interpro.

Interpro