Helix-Turn-Helix Motif Detection in Protein Sequences

Giri Narasimhan[1], Changsong Bu[2], Yuan Gao[3], Tom Milledge[1],
Xuning Wang[4], Ning Xu[5], Gaolin Zheng[1], And Kalai Mathee[5]

[1]School of Computer Science, Florida International University, [2]Idax Inc., [3]IBM T.J. Watson Research,
[4]Parke Davis, [5]University of Memphis, [6]Department of Biology, Florida International University.

ABSTRACT

We use methods from Data Mining and Knowledge Discovery to design an algorithm for detecting motifs in protein sequences. The algorithm assumes that a motif is constituted by the presence of a "good" combination of residues in appropriate locations of the motif. The algorithm attempts to compile such good combinations into a "pattern dictionary" by processing an aligned training set of protein sequences. The dictionary is subsequently used to detect motifs in new protein sequences. Statistical significance of the detection results are ensured by statistically determining the various parameters of the algorithm.

Based on this approach, we have implemented a program called GYM. The Helix-Turn-Helix (HTH) Motif was used as a model system on which to test our program. The program was also extended to detect Homeodomain motifs. The detection results for the two motifs compare favorably with existing programs.

Motifs in Protein Sequences

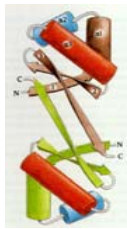
Motifs are combinations of secondary structures in proteins with a specific **structure** and a specific **function** Protein families are often characterized by one or more such motifs. Motif detection in proteins is thus an important problem since motifs carry out and regulate various functions, and the presence of specific motifs may help classify a protein.

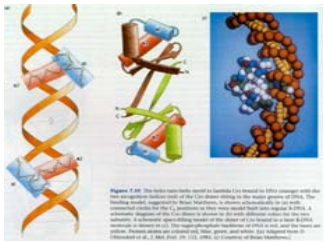
Examples: Helix-Turn-Helix, Zinc-finger, Homeobox domain, Hairpin-beta motif, Calcium-binding motif, Beta-alpha-beta motif, Coiled-coil motifs.

Helix-Turn-Helix Motifs

Structure

- 3-helix complex
- Length: 22 amino acids
- Turn angle





Function

Gene regulation by binding to DNA

Previous Methods

Profile Method: [Gribskov '90]

- Build a Profile matrix based on frequencies of occurrence of amino acids in specified locations within the motif. $\text{Weight}(i,AA) = 100 \log (p(i,AA) / (\pi(AA)\ N))$
- Use the profile to perform detection of the motif in new sequences.

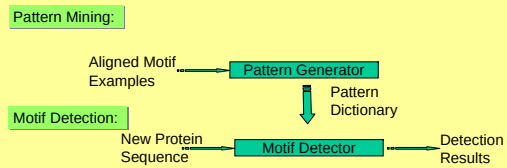
Hidden Markov Models

Neural Networks

New Algorithm: Basic Assumptions

- Combinations of residues in specific locations (may not be contiguous) contribute towards stabilizing a structure.
- Some reinforcing combinations are relatively rare.

New Algorithm: Outline



GYM Algorithm: Pattern Mining

Algorithm Pattern-Mining
Input: Motif length m, support threshold T,
list of aligned motifs M.
Output: Dictionary L of frequent patterns.

1. L_1 := All frequent patterns of length 1
2. **for** i = 2 **to** m **do**
3. C_i **:= Candidates**(L_{i-1})
4. L_i := Frequent candidates from C_i
5. **if** ($|L_i|$ <= 1) **then**
6. **return** L as the union of all L_j, j <= i.

GYM Algorithm: Motif Detection

Algorithm Motif-Detection

Input : Motif length m, threshold score T, pattern dictionary L,
and input protein sequence P[1..n].
Output : Information about motif(s) detected.

1. **for** each location i **do**
2. S := **MatchScore**(P[i..i+m-1], L).
3. **if** (S > T) **then**
4. Report it as a possible motif

Patterns: Examples

Loc	Protein	Helix 2						Turn				Helix 3							
	Name	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
14	Cro	F	G	Q	E	K	T	A	K	D	L	G	V	Y	Q	S	A	I	N	K	A	I	H
16	434 Cro	M	T	Q	T	E	L	A	T	K	A	G	V	K	Q	Q	S	I	Q	L	I	E	A
11	P22 Cro	G	T	Q	R	A	V	A	K	A	L	G	I	S	D	A	A	V	S	Q	W	K	E
31	Rep	L	S	Q	E	S	V	A	D	K	M	G	M	G	Q	S	G	V	G	A	L	F	N
16	434 Rep	L	N	Q	A	E	L	A	Q	K	V	G	T	T	Q	Q	S	I	E	Q	L	E	N
19	P22 Rep	I	R	Q	A	A	L	S	K	M	V	G	V	S	N	V	A	I	S	Q	W	E	R
24	CII	L	G	T	E	K	T	A	E	A	V	G	V	D	K	S	Q	I	S	R	W	K	R
4	LacR	V	T	L	Y	D	V	A	E	Y	A	G	V	S	Y	Q	T	V	S	R	V	V	N
167	CAP	I	T	R	Q	E	I	S	Q	I	V	G	C	S	R	E	T	V	G	R	I	L	K
66	TrpR	M	S	Q	R	E	L	K	N	E	L	G	A	G	I	A	T	I	T	R	G	S	N
22	BlaA Pv	L	N	F	T	K	A	A	L	E	L	Y	V	T	Q	G	A	V	S	Q	Q	V	R
23	TrpI Ps	N	S	V	S	Q	A	A	E	Q	L	H	V	T	H	G	A	V	S	R	Q	L	K

- Q1 G9 N20
- A5 G9 V10 I15

Experimental Results

Motif	Protein Family	Number Tested	GYM = DE Agree	Number Annotated	GYM = Annot.
HTH Motif (22)	Master	88	88 (100 %)	13	13
	Sigma	314	284 + 23 (98 %)	96	82
	Negates	93	86 (92 %)	0	0
	LysR	130	127 (98 %)	95	93
	AraC	68	57 (84 %)	41	34
	Rreg	116	99 (85 %)	57	46
	Total	675	653 + 23 (94 %)	289	255 (88 %)

Conclusions

- Pattern Discovery is a powerful way to detect motifs in protein sequences.
- A Pattern Dictionary is a composite descriptor of a motif or a domain, and is better than a consensus sequence or a motif signature.
- The GYM program is accurate, sensitive, and efficient. It also provides lot of useful information along with the motif detection.
- Identical patterns occurring in different proteins have been observed to often share near-identical structures.
- Pattern discovery can also be used as a basis for local and global structure prediction

Future Work

- Automate Motif Detection Process for other targeted motifs.
- Automating the Choice of a Training Set.
- Negative Training Set.
- Mutational Studies.
- Protein Structure Prediction.
- Functional Annotations.

Website for GYM Online: **www.cs.fiu.edu/~giri/bioinf/GYM2/welcome.html**