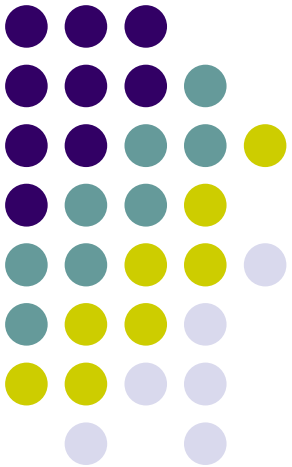
Varun: Extraction of Over-represented Extensible Motifs

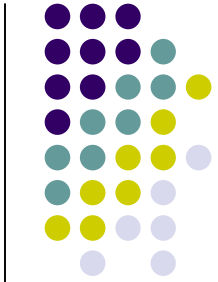
Alberto Apostolico
Purdue Univ. & Univ. of Padova
axa@dei.unipd.it

Matteo Comin
University of Padova
ciompin@dei.unipd.it

Laxmi Parida
IBM T.J. Watson Center
parida@us.ibm.com



Outline



- Biological Motivation
- A new Syntactic characterization: Extensible Motifs
- Evaluating Surprisingness (Probabilistic)
- Shrinking the search space by Monotonicity, Syntactical and Statistical properties
- Discovering Over-represented Extensible Motifs
- Algorithmic details
- Experiments and Validation
- Conclusion and Future work

Biological Motivations

- Regularity/Pattern/Motif are widely used to characterized Biological Data
- Everytime an alignment includes insertion or deletion it can not be described by a rigid pattern
- The Need of an expressive description without an exhaustive search

An Example from Prosite

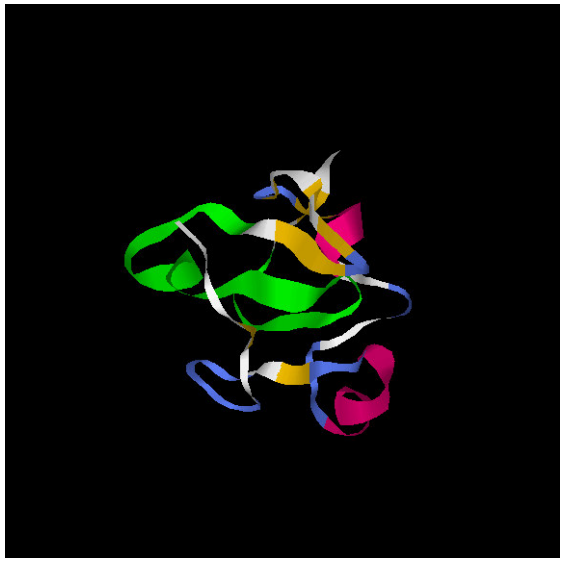
Entry name: HIPIP

Accession number: PS00596

Description: High potential iron-sulfur proteins signature.

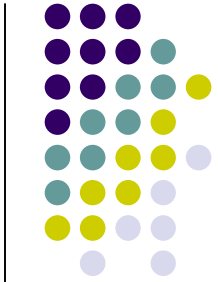
Pattern: C-(6,9)[LIVM]…G[YW]C..[FYW]

PDB 1PIJ



PDB 1HLQ

Modeling Patterns



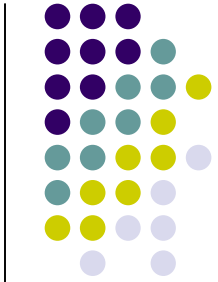
- Probabilistic models

  (e.g. Weight Matrices, Profiles)
- Deterministic models

  (e.g. words, motifs with don't cares)

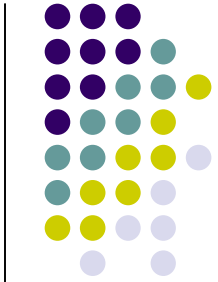
Brazma, Jonassen, Eidhammer, Gilbert
Approaches to the automatic discovery of patterns in biosequences, JCB 1998.

Previous Work



- Syntactic :
  - Teiresias
  - Irredundant Motifs
  - PRATT
- Statistical :
  - Gibbs sampler
  - Verbumculus
  - Splash
  - MEME

Extensible Motifs



Definition: *Extensible Motifs* are patterns which allow variable-length don't cares

e.g., Prosite **F…..G-(2,4)G.H**

- Note that the length of these patterns is variable
- High expressive power
- Huge pattern space

Extensible Motifs

(Implications of Variable-Gaps)

s = axbcaxxbcaxxxbc

m = a-[1-3]bc at pos 1, 5 and 10

Main Issues

1) a location list corresponds to multiple patterns

Eg. axbcpdaycbqd (at positions 1 and 7)

m1 = a-[1-2]b-[1-2]d

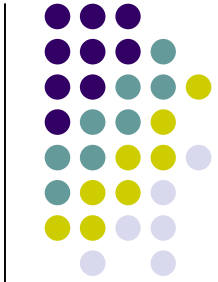
m2 = a-[1-2]c-[1-2]d

2) multiple occurrences at a location

Eg. axbbxc (at position 1)

m = a-[1-2]b-[1-2]c

Definitions (extensible patterns)



Realization of m m = a-[2-3]b m' = a..b m'' = a...b

$m_1 < m_2$ for every realization m'_1 of m_1, there exists a realization m'_2 of m_2 s.t. m'_1 is contained in m'_2

Maximality

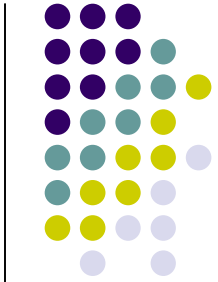
- in composition
- in length
- in extension

Over-represented Extensible Motifs

Two basic hypotheses about the source (probabilistic models)

- *Bernoulli*: symbols are generated independently and they are identically distributed (*i.i.d.* or *memoryless*)
- *Markov*: the probability distribution for the "next" symbol depends on the previous *h* symbols (*h>0* is the *order* of Markov chain)

Evaluating Over-representation



Z-Score-like functions :

$$z(w) = \frac{f(w) - E(w)}{N(w)}$$

f(w) : is the frequency of w

E(w) : the expected frequency of w

N(w) : represent the expected value of some function of w

Exploring Multiple Realizations

Let m be a realization of m, under the i.i.d

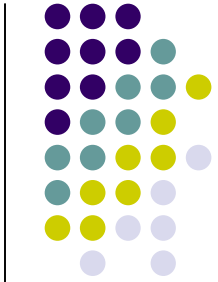
$$p_{\overline{m}} = \prod_{\sigma \in \Sigma} (p_\sigma)^{j_\sigma}$$

A motif is *degenerate* if it can possibly have multiple occurrences at a site.

LEMMA 1. *Let m be an extensible non-degenerate motif generated by a stationary,* iid *source which emits $(\sigma \in \Sigma)$ with probability p_σ. Let j_σ be the number of times σ appears in m and let e be the number of annotated dots in m with annotations $\alpha_1, \alpha_2, \ldots \alpha_e$. Then*

$$p_m = \prod_{\sigma \in \Sigma} (p_\sigma)^{j_\sigma} \prod_{i=1}^{e} |\alpha_i| \qquad (2)$$

Exploring Multiple Realizations



Let M^s denote a set of strings that has only the solid characters of at least *s* realizations of *m*.

Example:

m= *a-[1-3]b* with realizations *a.b*, *a..b* and *a...b*

M^1={*a.b*, *a..b*, *a...b*} M^2={*a.bb*, *a..bb*, *a.b.b*} M^3={*a.bbb*}

COROLLARY 2. *Let m be a degenerate (possibly with multiple occurrences at a site) extensible motif, and let* $p_{m^k} = \sum_{m' \in M^{k+1}} p_{m'}$; *then*

$$p_m = \sum_{k=0}^{r-1} (-1)^k (p_{m^{k+1}}). \qquad (4)$$

Monotonicity of Surprisingness

THEOREM 1. *Let v and u be possibly degenerate extensible motifs under the iid model and let v be a condensation of u. Then, there is an integer $\hat{p} \leq 1$ such that $p_v = p_u \hat{p}$.*

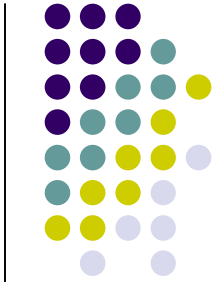
THEOREM 2. *If $f(u) = f(v) > 0$, $N(v) < N(u)$, and $E(v)/N(wv) \leq E(u)/N(u)$, then*

$$\frac{f(v) - E(v)}{N(v)} > \frac{f(u) - E(u)}{N(u)}$$

THEOREM 3. *Let u and v be motifs generated with respective probabilities p_u and $p_v = p_u \hat{p}$ according to an iid process. If $f(u) = f(v)$ and $p_u < 1/2$ then*

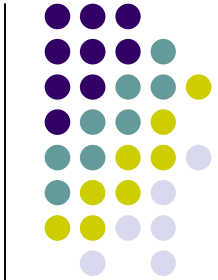
$$\frac{f(v) - E(v)}{\sqrt{E(v)(1 - p_v)}} > \frac{f(u) - E(u)}{\sqrt{E(u)(1 - p_u)}}$$

Lessons



- A non-maximal motif is a submotif or a condensation of a maximal one which is no less surprising.
- It is enough extract and score only maximal motifs.

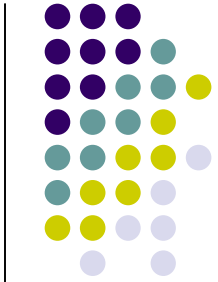
Discovering Extensible Motifs



```
Main()
{
Result ← {};
B ← {m_i | m_i is a cell};
For each m = Extract(B)
    Iterate(m, B, Result);
Result ← Result;
}
```

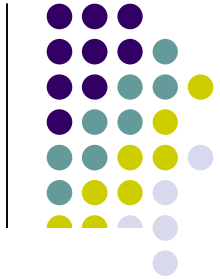
```
Iterate(m, B, Result)
{
        m' ← m;
        For each b = Extract(B) with
         ((b ∼–compatible m')
           OR (m' ∼–compatible b))
              If (m' ∼–compatible b)
                 m_t ← m' ∼ b;
                 If NodeInconsistent(m_i) exit;
                 If (|L_m'| = |L_b|) B ← B − {b};
                 If (|L_m'| ≥ K)
                      m' ← m_t;
                        Iterate(m', B, Result);
              If (b ∼–compatible m')
                 m_t ← b ∼ m';
                 If NodeInconsistent(m_i) exit;
                 If (|L_m'| = |L_b|) B ← B − {b};
                 If (|L_m'| ≥ K)
                      m' ← m_t;
                      Iterate(m', B, Result);
              For each r ∈ Result with Z_r = Z_m'
                 If (m' is not maximal w.r.t. r)
                      return;
              Result ← Result ∪ {m'};
}
```

Different Pruning Criterias



- Combinatorial pruning
  1. Pruning by Occurrences:
     - -k<Num>: Num is the quorum or the minimum number of times a pattern must occur in the input.
     - -c: When this is specified the quorum k is in terms of the number of sequences where the pattern occurs at least once.
  2. Pruning by composition:
     - Using homology groups: -b<File>: File lists the symbol equivalences that define the homology groups.
     - -R: When this mode is specified, only rigid patterns are discovered.
     - Extensibility:
       - -D<Num>: Num is the maximum number of consecutive don't care characters ('.') in the realization of an extensible pattern.
       - -d<Num>: Num is the minimum number of nonextensible characters (including the don't care character) between two consecutive extensible characters
- Statistical pruning
  1. -p<File>: File lists the symbol probabilities used for the probabilistic analysis.
  2. -z<Val>: Val is the minimum absolute value of Zscore of the patterns.

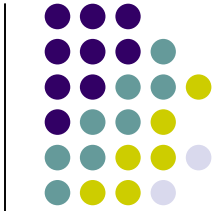
Experiments



Rank	z-score	Motif
1	**1497,62**	**C-(6,7,8,9)[LIVM]...G[YW]C..[FYW]**
2	978,872	P-(3,4,6,8,9)[LIVM]...G[YW]C..[FYW]
3	590,866	C-(6,7,8,9)[LIVM]...G[YW]C-(1,3,4,5,6,7)A
4	564,821	C-(6,7,8,9)[LIVM]...G[YW]C-(1,3,4,5,6,7)[ATD]
5	537,73	[LIVM]-(1,2,3,4,5,7,8,9)G[YW]C..[FYW]
6	385,2	[LIVM]-(1,2,3,4,5,7,8,9)G[FYW]C..[FYW]
7	161,173	[LIVM]...G[FYW]C-(2,4)[FYW]
8	156,184	[LIVM]-(1,2,3,4,5,6,7,8,9)G[YW]C
9	138,881	[LIVM]-(1,3,4,5,6)[LIVM]...G[FYW]C-(1,3,4,5,6,7)A

Fig. 2. The functionally relevant motif is shown in bold for high potential iron-sulfur proteins (HiPIP) (id PS00596), Here 22 sequences of about 2500 bases were analyzed at k=22, D=9, d=4.

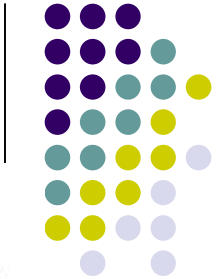
Experiments



Rank	z-score	Motif
1	**295840**	**[LIM]-(1,2,3,4)[STA][FY]DPC[LIM][ASG]C[ASG].H**
2	**2,86E+05**	**[LIM]-(1,2,3,4)[ASG][FY]DPC[LIM][ASG]C[ASG].H**
3	**155736**	**R-(1,4)[FY]DPC[LIM][ASG]C[ASG].H**
4	78829	[LIM]-(1,2,3,4)[STA].DPC[LIM][ASG]C[ASG].H
5	76101,9	[LIM]-(1,2,3,4)[ASG].DPC[LIM][ASG]C[ASG].H
6	34205,6	[STA]-(1,4)DPC[LIM][ASG]C[ASG].H
7	30325,1	[LIM]-(1,2,3,4)[STA][FY]D.C[LIM][ASG]C..H
8	29276	[LIM]-(1,2,3,4)[ASG][FY]D.C[LIM][ASG]C..H
9	20527,3	[ASG]-(1,4)DPC[LIM][ASG]C[ASG].H
10	17503,4	[LIM]-(1,2,3,4)[ASG]..PC[LIM][ASG]C[ASG].H

Fig. 4. The functionally relevant motifs are shown in bold for Nickel-dependent hydrogenases (id PS00508). Here 22 sequences of about 23000 bases were analyzed at k=22, D=4, d=3.

Experiments



Rank	z-score	Motif
1	**7,60E+07**	**RA.T[LV].C.P-(2,3)G.HP....AC[ATD].L....[ASG]**
2	21416,8	A..[LV].C.P-(2,3)G.HP-(1,2,4)[ASG].[ATD]
3	8105,33	A-(1,4)T....P-(2,3)G.HP....[ATD]-(3)L....[ASG]
4	5841,85	[ATD].T....P-(1,2,3)G.HP-(1,2,4)A.[ATD]
5	4707,62	P.[ASG]-(2,3,4)P....AC[ATD].L....[ASG]
6	4409,21	A..[LV]...P-(2,3)G.HP-(1,2,4)A.[ATD]
7	3086,17	P-(1,2,3)[ASG]..P-(4)AC[ATD].L....[ASG]
8	3068,18	R..[ATD]....P-(2,3)G.HP-(1,2,4)[ASG].[ATD]
9	2615,98	[ASG][ATD]-(1,3,4)P....AC[ATD].L....[ASG]
10	2569,66	[ASG]-(1,2,3,4)P....AC[ATD].L....[ASG]
11	2145,6	G-(2,3)P....AC[ATD].L....[ASG]

Fig. 3. The functionally relevant motif is shown in bold for Streptomyces subtilisin-type inhibitors signature (id PS00999). Here 20 sequences of about 2500 bases were analyed at k=20, D=4, d=4.

Rank	z-score	Motif
1	2,84E+09	Y...L...C..[FYW]A..[STAH]R..P..FNE[STAH]K.I.F[STAH]M
2	8,28E+07	V-(1,3,4)G...S..[STAH]....N...L....Q-(4)[STAH]....L.[DN]...[FYW]..F....P....Q..A...I
3	5,55E+07	L-(2,3)F...Q....[STAH][STAH]...L.[DN]...[FYW]..F.R..P.D..Q..A...I
4	4,27E+07	L-(2,3)F...Q.[STAH]..[STAH][STAH]....S....[FYW]..F.R..P.D..Q..A...I
5	4,23E+07	L....I...[STAH]..[STAH]....LS[DN]...[FYW]..F.R..P.D..Q..A...I
6	3,99E+07	LF-(3)Q....[STAH][STAH]....S[DN]...[FYW]..F.R..P.D..Q..A...I
7	3,38E+07	LF-(3)Q....[STAH][STAH]...L.[DN]...[FYW]..F.R..P.D..Q..A...I
8	3,38E+07	LF...Q....[STAH]-(4)L.[DN]...[FYW]..F.R..P.D..Q[STAH].A...I
9	3,29E+07	I-(1)Q.[STAH]..[STAH]....LS[DN]...[FYW]..F.R..P.D..Q..A...I
10	3,29E+07	I.Q-(4)[STAH]....LS[DN]...[FYW]..F.R..P.D..Q[STAH].A...I
11	3,29E+07	I.Q.[STAH]..[STAH]-(4)LS[DN]...[FYW]..F.R..P.D..Q..A...I
12	3,10E+07	L....Q-(1,4)[STAH]..[STAH]....LS[DN]...[FYW]..F.R..P.D..Q..A...I
13	2,77E+07	L[FYW]-(3)Q.[STAH]..[STAH]....LS....[FYW]..F.R..P.D..Q..A...I
14	2,58E+07	L-(4)Q.[STAH]..[STAH]....LS[DN]...[FYW]..F.R..P.D..Q..A...I
15	2,30E+07	S.[STAH]S-(2,4)LS[DN]...[FYW]..F.R..P.D..Q[STAH].A...I
16	2,15E+07	L-(1,3,4)C..[FYW]A..[STAH]R..P..F.E.K.I.F.M
17	1,40E+07	F-(1)I.Q...[STAH][STAH]-(4)L[STAH]....[FYW]..F.R..P.D..Q..A...I
18	1,37E+07	L-(2,4)I...[STAH].[STAH].[STAH]-(3)LS....[FYW]..F.R..P.D..Q..A...I
19	1,02E+07	L..I-(1)Q....[STAH][STAH].....S....[FYW]..F.R..P.D..Q..A...I
20	8,65E+06	I-(1)Q....[STAH][STAH]...L.[DN]...[FYW]..F.R..P.D..Q..A...I
21	8,19E+06	S[STAH]-(1,2,3,4)LS[DN]...[FYW]..F.R..P.D..Q[STAH].A...I
22	7,98E+06	Q-(3)[STAH][STAH]....LS[DN]...[FYW]..F.R..P.D..Q..A...I
23	6,82E+06	F-(3)Q....[STAH][STAH]...L[STAH]....[FYW]..F.R..P.D..Q..A...I
24	5,66E+06	A[STAH][STAH]-(2,3)LS[DN]...[FYW]..F.R..P.D..Q..A...I
25	5,57E+06	F.I-(3)[STAH]..[STAH]....L[STAH]....[FYW]..F.R..P.D..Q..A...I
26	5,18E+06	L.L-(4)Q....[STAH]....L-(1)[DN]...[FYW]..F.R..P.D..Q..A...I
27	3,61E+06	L.L-(2)I...[STAH]....[STAH]....[STAH]....[FYW]..F.R..P.D..Q..A...I
28	3,48E+06	[STAH].[STAH]-(1,2,3)LS[DN]...[FYW]..F.R..P.D..Q..A...I
29	3,17E+06	[STAH]...[STAH]...LS[DN]...[FYW]..F.R..P.D..Q..A...I
30	2,47E+06	L....Q-(4)[STAH][STAH]....S....[FYW]..F.R..P.D..Q..A...I
31	2,43E+06	V-(1,3)N.L....I-(3)[STAH]...[STAH]....[STAH]....[FYW]..F....P.D..Q..A...I
32	2,22E+06	[STAH][STAH][STAH]-(1,2,3)LS....[FYW]..F.R..P.D..Q..A...I
33	2,06E+06	[STAH].[STAH][STAH]....LS....[FYW]..F.R..P.D..Q..A...I
34	2,03E+06	Y...L...C....A...R..P..F.E.K.I-(1,4)[FYW][STAH]
35	1,99E+06	I.Q...[STAH]-(1)[STAH]....L.[DN]...[FYW]..F....P.D..Q..A...I
36	1,99E+06	I.Q-(1)[STAH]...[STAH]....L.[DN]...[FYW]..F....P.D..Q..A...I
38	1,97E+06	F.I...[STAH]-(3)[STAH]....L.[DN]...[FYW]..F....P.D..Q..A...I
40	1,97E+06	F.I-(3)[STAH]..[STAH]....L.[DN]...[FYW]..F....P.D..Q..A...I
41	1,91E+06	[STAH]..[STAH].K-(1,4)P..FNE[STAH]K.I.F[STAH]M
42	**1,72E+06**	**CC[FYW].C..C....[FYW]-(2,4)[DN]..[STAH]C..C**
43	1,57E+06	[STAH]-(1,3,4)[FYW]A..[STAH]R..P..F.E.K.I.F.M
44	1,49E+06	A-(1,3)[STAH]...L[STAH][DN]...[FYW]..F.R..P.D..Q..A...I
45	1,36E+06	Q...[STAH].[STAH]-(3)L[STAH]....[FYW]..F.R..P.D..Q..A...I
46	1,32E+06	I-(3)[STAH]..[STAH][STAH].....S....[FYW]..F.R..P.D..Q..A...I
47	1,31E+06	[STAH][STAH]-(1,2,3,4)L.[DN]...[FYW]..F.R..P.D..Q..A...I
48	1,24E+06	[STAH]..[STAH][STAH]-(1,3)LS....[FYW]..F.R..P.D..Q..A...I
49	1,19E+06	[FYW]-(1,3,4)[STAH]...P..FNE[STAH]K.I.F[STAH]M
50	1,12E+06	I...[STAH]-(3)[STAH]...L[STAH]....[FYW]..F.R..P.D..Q..A...I

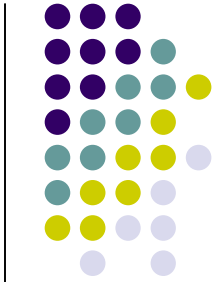
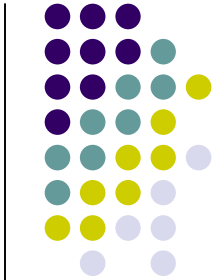


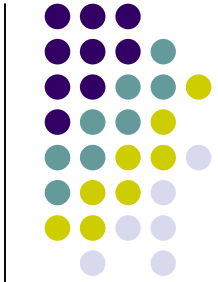
Fig. 5. The functionally relevant motif is shown in bold for G-protein coupled receptors family 3 (id PS00980). This run involved 25 sequences of about 25000 bases each at k=25, D=4, d=8.

Conclusion



- A notion of Extensible Motif proves useful in biosequence analysis
- Maximality and monotonicity bound search space
- An efficient algorithm detects over-represented extensible motifs
- Results shows good sensitivity and selectivity
- Interesting applications have emerged also to other contexts: data compression

Future Work



- Speed-up the Algorithm or prove a tigther bound
- Extend the probability models
- Availability (?/?/05):

http://www.research.ibm.com/bioinformatics/

Refences

- "Conservative Extraction of Over-represented Extensible Motifs", A. Apostolico, M. Comin, L. Parida, To appear in 2005 ISMB.
- “An inexact suffix tree based algorithm for extensible pattern discovery”, Chattaraj,A. and Parida,L. *Theoret. Comput. Sci.*, **335**: 3–14 (2005).
- "Bridging Lossy and Lossless Compression by Motif Pattern Discovery", M. Comin, A. Apostolico, L. Pardia, *General Theory of Information Transfer and Combinatorics*, Vol. II, 2004, edit by R. Ahlswede.
- “Mining, Compressing and Classifying with Extensible Motifs", A. Apostolico, M. Comin, L. Parida, under submission.

Acknowledgements

- ISMB Travel fellowship
- IBM Internship support
- Italian Ministry of University and Research and the Research Program of the University of Padova