Instructor: Sael Lee

CS549 Spring – Computational Biology

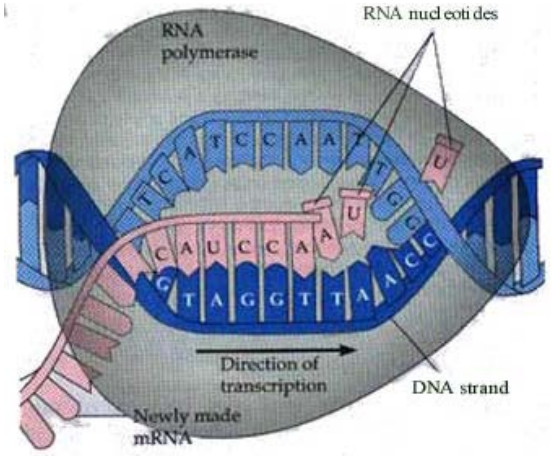
LECTURE 4:
DNA BINDING AND INFORMATION THEORY

A BRIEF REVIEW OF MOLECULAR INFORMATION THEORY.

SCHNEIDER, T. D. , (2010). *NANO COMMUNICATION NETWORKS1*(3), 173–180.

MOLECULAR INFORMATION THEORY

- **Molecular information theory:** Using information theory to measure states and patterns of molecules.
- Problem we focus on: **Interaction between DNA and Protein**

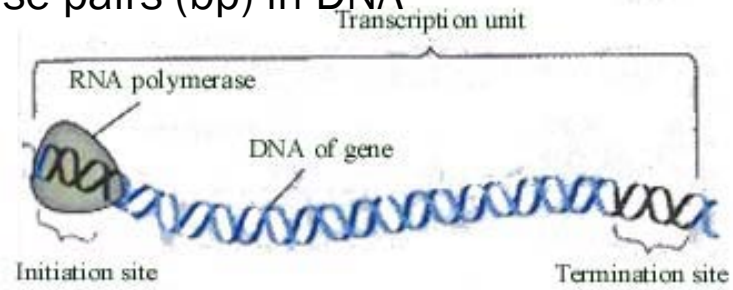


Transcription process:
RNA Polymerase (protein) binding to DNA

PROBLEM:
Analysis of interaction between DNA and proteins that control the expression DNA

PROPERTIES:

- Protein is a finite molecule
- Interaction content of proteins cover 10-20 base pairs (bp) in DNA

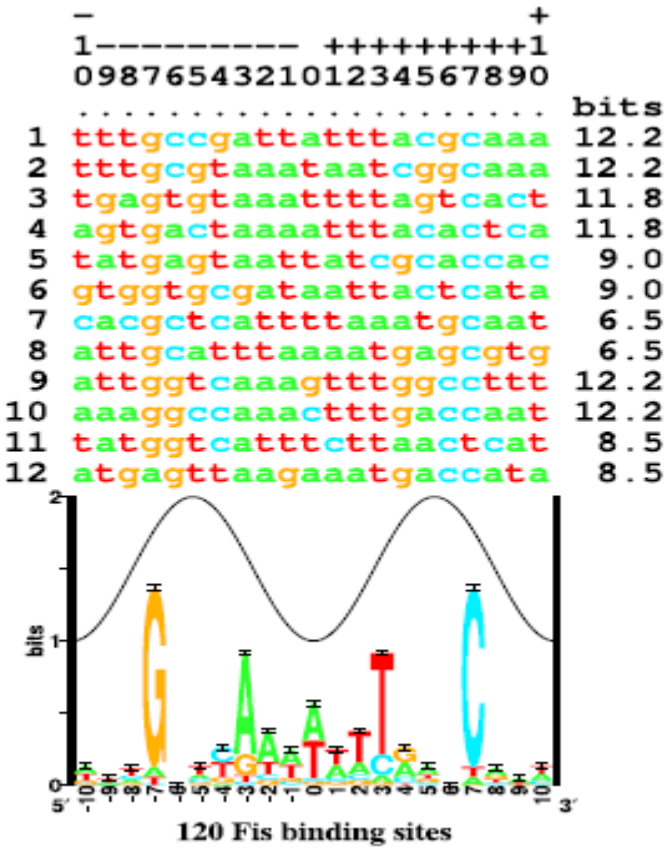


Interaction site: 10~20 bp

SEQUENCE LOGO – REVIEWED

- **Sequence logo** is a graphical representation of the sequence **conservation** of nucleotides (in a strand of DNA/RNA) or amino acids (in protein sequences)
- They can show how much pattern is in a set of binding sits.

Schneider & Stephens (1990) NAR. 18: 6097-6100



EX> Fis site

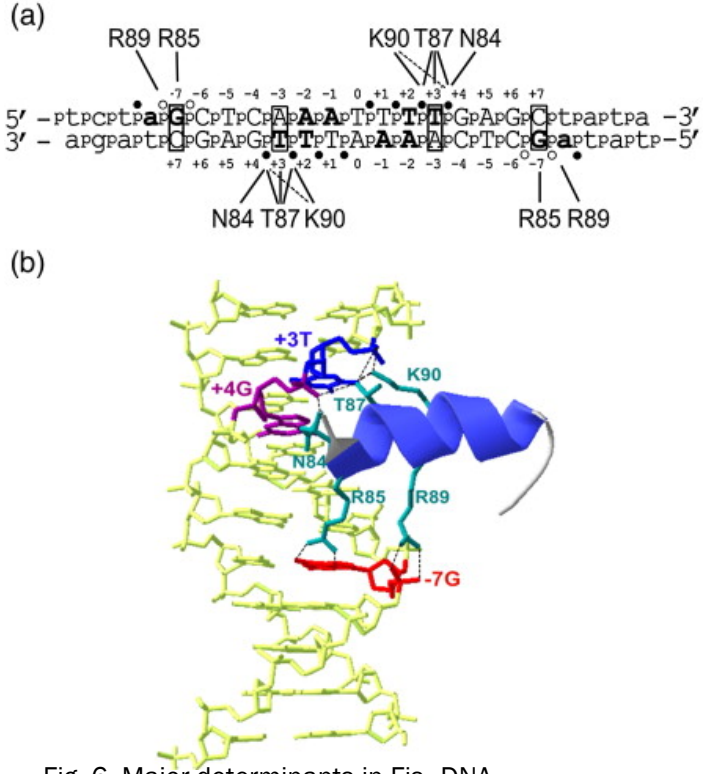


Fig. 6. Major determinants in Fis–DNA binding. [Shao et al. (2008) JMB 380:2, 327-339.]

SEQUENCE LOGO

The **information content** (y-axis) of position *i* :

Four letter: A,T,C,G Entropy small-sample correction

$$R_i = \log_2(4) - (H_i + e_n)$$

$Entropy\ H_i$ computed as

$$H_i = -\sum_{a \in \{A,T,C,G/U\}} f_{a,i} * \log f_{a,i}$$

Where $f_{a,i}$ is relative frequency of bases *a* at position *I* and
e_n small-sample correction for an alignment of *n (4 for DNA/RNA)* letters

$$e_n = \frac{1}{ln2} * \frac{s-1}{2n}$$

where s is 4 for nucleotides, 20 for amino acids, and n is the number of sequences in the alignment.

- The total height of the letters depicts the information content of the position, in bits

The height of letter *a* in column *i* is given by

$$\boldsymbol{height_a = f_{a,i} * R_i}$$

CHARACTERIZING BINDING SITES

- Before binding, protein is uncertain as to what base it will see and that uncertainty can be measured as $\log_2(4)$
  - Before we know the binding event can occur, all four bases (A,T,C,G) can be seen in a DNA locus.
- After binding, uncertainty of what it is touching in different cases is lower.
  - If only one type of bases occur:

$$\log_2(1) = 0$$

  - If other bases occur as well: (Conditional Entropy)

$$H(i) > 0$$

The **information content** (y-axis) of position *l*:

Height in sequence logo

Four letter: A,T,C,G Entropy small-sample correction

$$R_{sequence}(i) = \log_2(4) - (H(i) + e_n) \quad \text{(bits per base)}$$

$$\mathrm{I}(X;Y) = H(X) - H(X|Y)$$

$\log_2(4)$: Uncertainty 'observed' by the DNA binding protein before binding to a site.

-> * maximum uncertainty possible: $\log_2 |\chi|$

$H(l)$: Uncertainty 'observed' by the DNA binding protein after binding to a site.

$$H(i) = -\sum_{b \in \{A,T,G,C\}} f_{b,i} \log_2 f_{b,i} \quad \text{(bits per base)}$$

where $f_{b,i}$ are the frequency of base b at a position i.

Assuming independence between sites,
total information in a binding site.

$$R_{sequence} = \sum_{l} R_{sequence}(l)$$

INFORMATION REQUIRED TO FIND A SET OF BINDING SITES

G = # of potential binding sites
= genome size in some cases
γ = number of binding sites on genome

Information required to find binding sites

Uncertainty before being bound to one of the sites

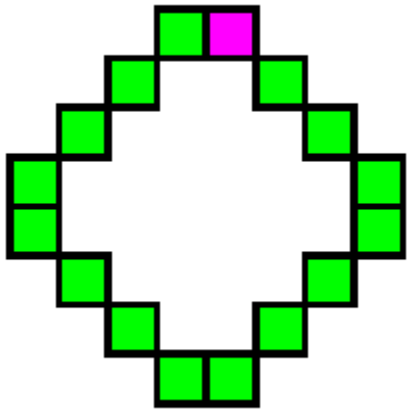
Uncertainty after being bound to one of the sites

$$R_{frequency} = H_{before\ binding} - H_{after\ binding}$$

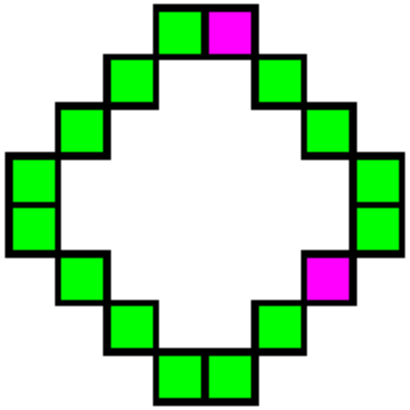
$$= \log_2 G - \log_2 \gamma$$

$$= -\log_2 \frac{\gamma}{G}$$ (bit per site)

INFORMATION REQUIRED TO FIND A SET OF BINDING SITES IN A GENOME



16 positions
1 site
$\log_2 16/1 = 4$ bits



16 positions
2 sites
$\log_2 16/2 = 3$ bits

Hypothesis:

The information in binding site patterns is just sufficient for the sites to be found in the genome

Natural Binding sites have $\boldsymbol{R_{sequence}}$ **closes to** $\boldsymbol{R_{frequency}}$

The information in the binding site pattern $\boldsymbol{R_{sequence}}$
is close to
The information needed to find the binding sites $\boldsymbol{R_{frequency}}$

But for a species in **a stable environment:**

- size of genome (G) is fixed (e. g. E. coli has 4.7 × 106 bp)
- number of binding sites (γ) is fixed (e. g. there are 50 E. coli LexA sites)

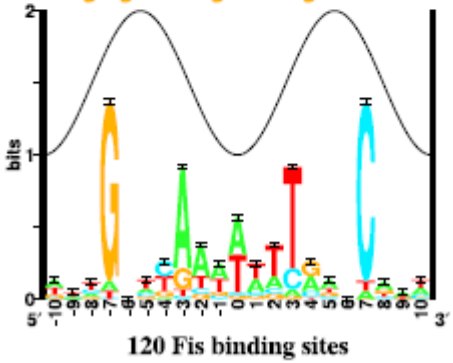
$$R_{frequency} = -\log_2 \frac{\gamma}{G}$$ is constant

$R_{sequence}$ must evolve towards $R_{frequency}$!

Made sense with simulated data

SEQUENCE WALKERS SHOW INDIVIDUAL INFORMATION OF BINDING SITES

```
   -                    +
   1--------- +++++++++1
   0987654321012345678 90
   ..................... bits
 1 tttgccgattatttacgcaaa 12.2
 2 tttgcgtaaataatcggcaaa 12.2
 3 tgagtgtaaattttagtcact 11.8
 4 agtgactaaaatttacactca 11.8
 5 tatgagtaattatcgcaccac  9.0
 6 gtggtgcgataattactcata  9.0
 7 cacgctcattttaaatgcaat  6.5
 8 attgcatttaaaatgagcgtg  6.5
 9 attggtcaaagtttggccttt 12.2
10 aaaggccaaactttgaccaat 12.2
11 tatggtcatttcttaactcat  8.5
12 atgagttaagaaatgaccata  8.5
```



120 Fis binding sites

$$R_{sequence} = avg(\log_2 f_{b,l})$$

A REEXAMINATION OF INFORMATION THEORY-BASED METHODS FOR DNA-BINDING SITE IDENTIFICATION.

ERILL, I., & O'NEILL, M. C. (2009). *BMC BIOINFORMATICS*, *10*, 57.

Abstract

Background: Searching for transcription factor binding sites in genome sequences is still an open problem in bioinformatics. Despite substantial progress, search methods based on information theory remain a standard in the field, even though the full validity of their underlying assumptions has only been tested in artificial settings. Here we use newly available data on transcription factors from different bacterial genomes to make a more thorough assessment of information theory-based search methods.

Results: Our results reveal that conventional benchmarking against artificial sequence data leads frequently to overestimation of search efficiency. In addition, we find that sequence information by itself is often inadequate and therefore must be complemented by other cues, such as curvature, in real genomes. Furthermore, results on skewed genomes show that methods integrating skew information, such as *Relative Entropy*, are not effective because their assumptions may not hold in real genomes. The evidence suggests that binding sites tend to evolve towards genomic skew, rather than against it, and to maintain their information content through increased conservation. Based on these results, we identify several misconceptions on information theory as applied to binding sites, such as negative entropy, and we propose a revised paradigm to explain the observed results.

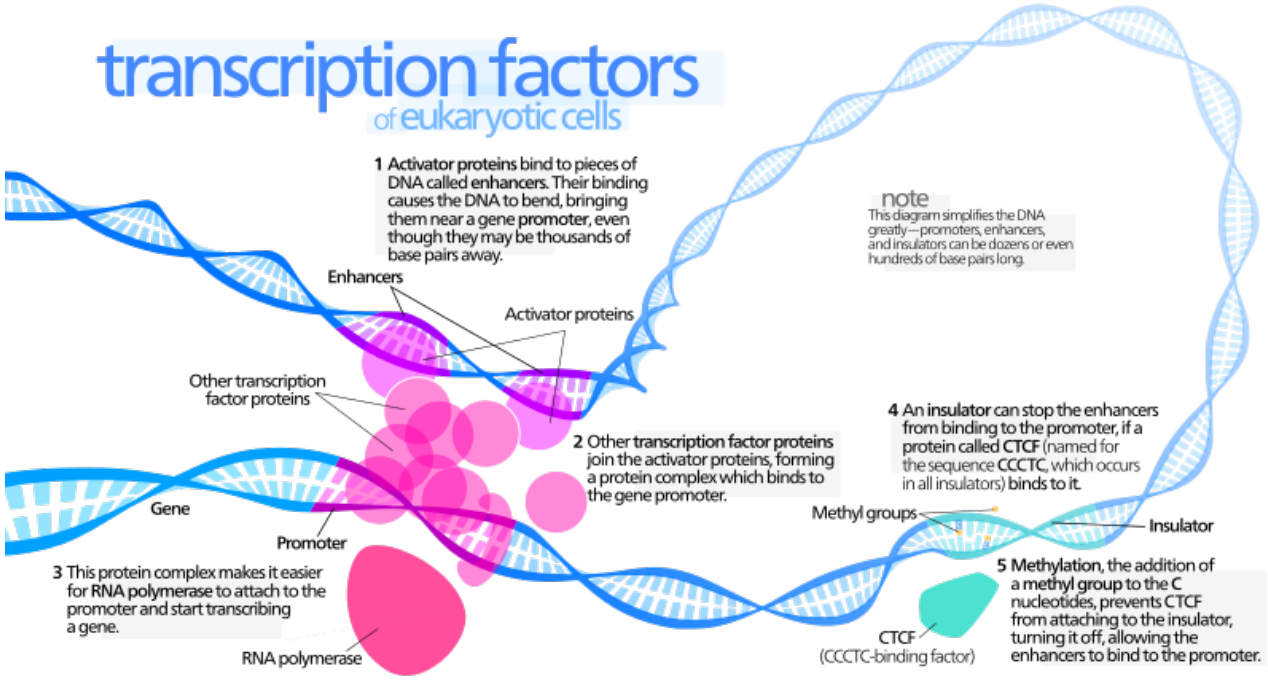
Conclusion: We conclude that, among information theory-based methods, the most unassuming search methods perform, on average, better than any other alternatives, since heuristic corrections to these methods are prone to fail when working on real data. A reexamination of information content in binding sites reveals that information content is a compound measure of search and binding affinity requirements, a fact that has important repercussions for our understanding of binding site evolution.

Conclusion

The results presented above have several important implications for the understanding of binding site search, information and evolution. On the search problem, we conclude that non-weighted $R_{sequence}$-based methods should be used preferentially, as they contain fewer assumptions and are thus less prone to misfire on real biological data. Conversely, weighted $R_{sequence}$-based methods seem to be better indicated to affinity rank sites. Relative entropy and similar heuristic corrections for skew composition should be avoided, since they are based on the misguided hypothesis that search and differential regulation are equivalent problems for the protein. In contrast, we propose that information content as defined by $R_{sequence}$ is a compound measure that incorporates requirements from the search and regulation processes. This revised paradigm suggests that binding sites will tend to drift towards the genomic skew, not against it, and increase their conservation to circumvent the global loss of information content in skewed genomes.

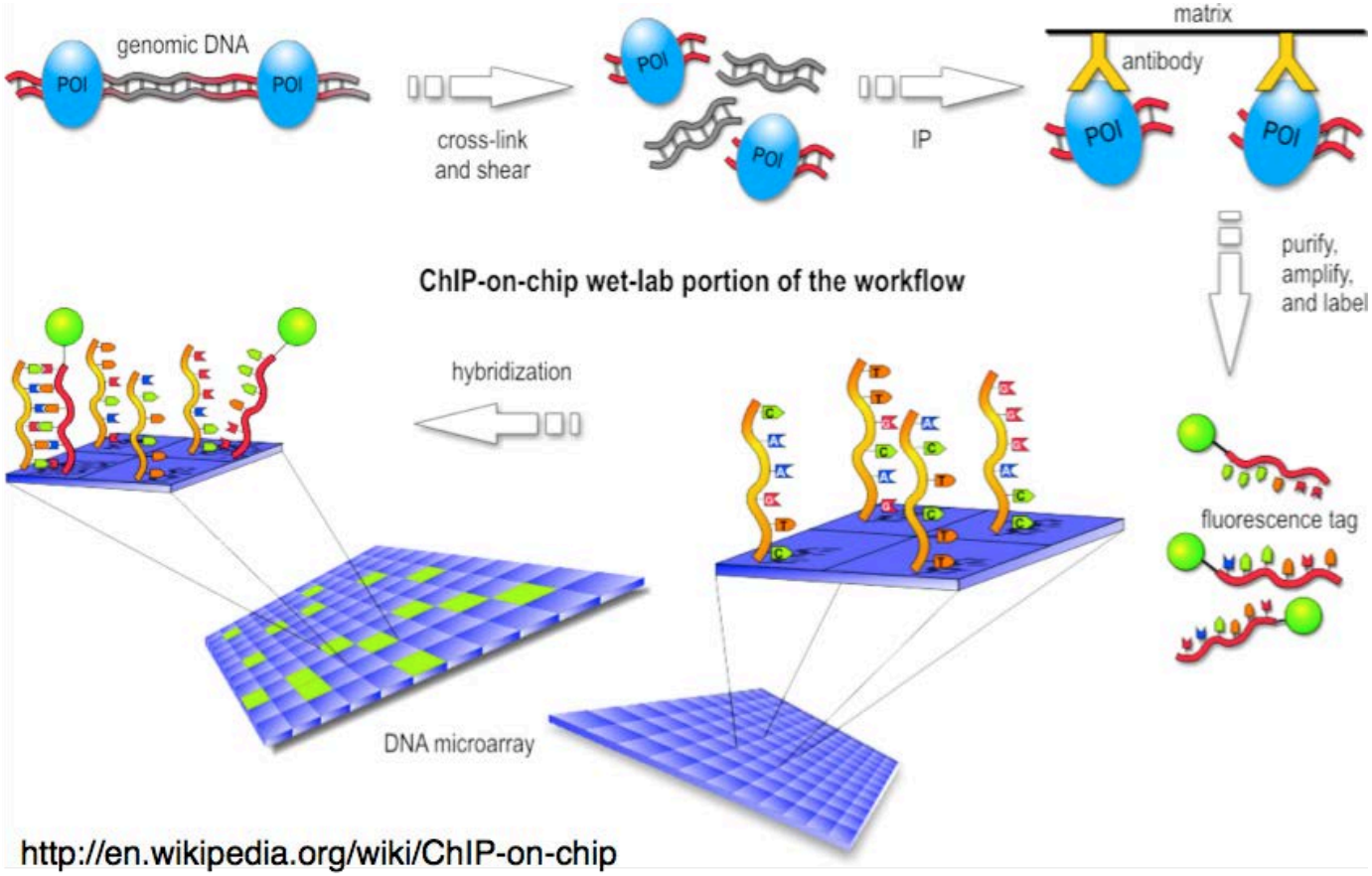
TRANSCRIPTION FACTOR (TF) BINDING

"In molecular biology and genetics, a **transcription factor** (sometimes called a sequence-specific DNA-binding factor) is a protein that binds to specific DNA sequences, thereby controlling the flow (or transcription) of genetic information from DNA to mRNA"



http://en.wikipedia.org/wiki/Transcription_factor

CHIP-CHIP EXPERIMENT



TWO PROBLEMS PROTEIN-DNA BINDING

- Affinity rank problem
  - Ranking which sequence will bind better
  - Measured in $R_{sequence}$:
- Site search problem
  - Finding the location of binding
  - Measured in $R_{frequency}$:
  - Assumes on/off binary affinity

Are they the same?

PREVIOUS ASSUMPTION

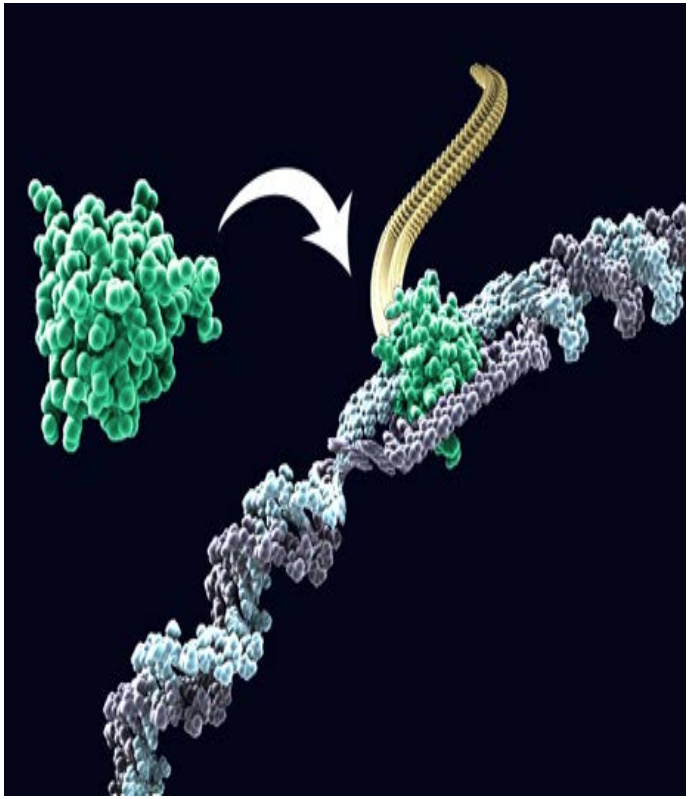
- Searching and ranking are the same binding problem

$R_{sequence}$ must evolve towards $R_{frequency}$!

WHAT THE DATA SAY

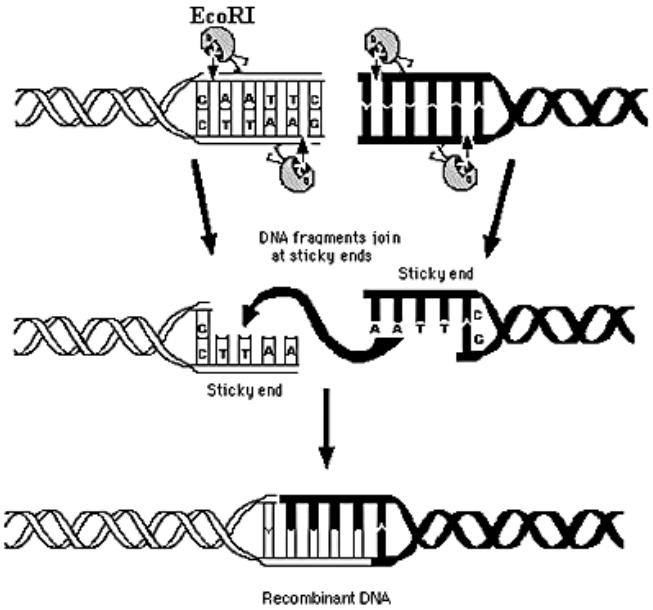
- Effective binding must be compound function of both the affinity of the protein from the site (ranking) and its ability to locate it within the genome (search)

Transcription factors



Background genome stays the same (H_{before}) while number of binding site change (H_{after})

restriction enzyme



Restriction Enzyme Action of EcoRI

Background genome change (H_{before}) while number of binding site stays the same (H_{after})

INFORMATION CONTENT OF A LOCATION

Information content accounting for uniform genome content

$$R_{sequence}(l) = -\sum_{S\in\{A,T,G,C\}}[f(S)\log_2 f(S)] - \left(-\sum_{S_l\in\{A,T,G,C\}}[p(S_l)\log_2 p(S_l)]\right)$$

$f(S)$: relative frequency in genome sequence

$p(S_l)$: frequency of each base S_l at position *l* in the prototype group

Information content accounting for skewed genome

$$RE(l) = R^{*}_{sequence}(l) = \sum_{S_l\in\Omega}\left(p(S^l)\cdot\log_2\left(\frac{p(S_l)}{f(S_l)}\right)\right)$$

$$RE(l) = \left[-\sum_{S_l\in\Omega}\left(\frac{p(S_l)}{f(S_l)}\cdot f(S_l)\cdot\log_2(f(S_l))\right)\right] - \left[-\sum_{S_l\in\Omega}(p(S_l)\cdot\log_2(p(S_l)))\right]$$

Weight

LIKELIHOOD THAT A SEQUENCE WAS A BINDING SITE FOR A GIVEN PROTEIN

Information content accounting for uniform genome content

information content of an individual binding sequence *i*

$$R_i(l) = \left[-\sum_{S \in \Omega} \left[f(S) \cdot \log_2 \left(f(S) \right) \right] \right] - \left[-\log_2 \left(p(S_{i,l}) \right) \right] = H_{before} - \left[-\log_2 \left(p(S_{i,l}) \right) \right]$$

Information content accounting for skewed genome

Explicitly takes into account the background genomic frequencies

$$I_i^{seq}(l) = p(S_{i,l}) \cdot \log_2 \left(\frac{p(S_{i,l})}{f(S_{i,l})} \right)$$

$Assumption$: $R_{sequence} \approx R_{frequency}$!

PUTTING IN RELATIVE IMPORTANCE OF EACH POSITION IN A MOTIF

- Accounting for importance of weight of each position in the prototype group
- Idea to make information in conserved region higher than the information of discordant region

$$R'_{sequence}(l) = R^{-}_{sequence}(l) \cdot \left(R^{+}_{sequence}(l) - R^{-}_{sequence}(l) \right)$$

calculated both before (-) and after (+) the addition of the query sequence to the prototype group.

- positive value:
  - query sequence concurs with the prototype since *R*+ will be improved by the addition,
- negative value:
  - query sequence discordant with the prototype

OBSERVATION 1

- Weighted vs non-weighted measures have different performance
  - Weighted better for binding the affinity rank
    - Conserved motif positions are the main players in determining the strength of a site
  - Non-weighted better for searching binding site
    - seems to be taking into account secondary information residing in poorly conserved positions that can be of relevance to the protein in order to make non-specific contacts or as a requirement for optimal curvature or bendability.
  - mean difference in search efficiency between weighted and non-weighted methods decreases as motif conservation increases

OBSERVATION 2

- $\left|R_{sequence} - R_{freqence}\right| > 0$
  - Ex> 20% of true CRP sites are left unaccounted for when using information theory-based methods for locating them.
- "Information lying in poorly conserved motif positions is being used actively by the protein to discern true binding sites against the genomic background."

- Experimental results have already hinted at the existence of several complementary sources of information for site location, such as curvature, pre-recruitment or cooperative binding .

- $R_{sequence}$: $R_{sequence}(l) = \log_2(4) - (H(l))$
  - uncertainty of the recognition process
- $R_{frequency}$: $R_{frequency} = -\log_2 \frac{\gamma}{G}$
  - uncertainty in terms of distinguishing a sequence from the genomic background

Conclusion

The results presented above have several important implications for the understanding of binding site search, information and evolution. On the search problem, we conclude that non-weighted $R_{sequence}$-based methods should be used preferentially, as they contain fewer assumptions and are thus less prone to misfire on real biological data. Conversely, weighted $R_{sequence}$-based methods seem to be better indicated to affinity rank sites. Relative entropy and similar heuristic corrections for skew composition should be avoided, since they are based on the misguided hypothesis that search and differential regulation are equivalent problems for the protein. In contrast, we propose that information content as defined by $R_{sequence}$ is a compound measure that incorporates requirements from the search and regulation processes. This revised paradigm suggests that binding sites will tend to drift towards the genomic skew, not against it, and increase their conservation to circumvent the global loss of information content in skewed genomes.

"This revised paradigm suggests that binding sites will tend to drift towards the genomic skew, not against it, and increase their conservation to circumvent the global loss of information content in skewed genomes."

Just means that if the sequence is more conserved, affinity probability is higher, and requires less additional factors for the protein recognize the binding sequence and *vise versa.*