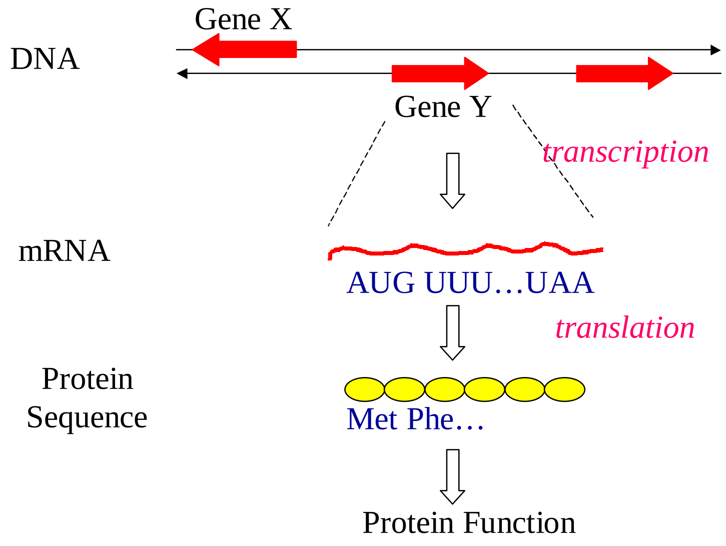
Introduction to Bioinformatics

KITP, UCSB
Hao Li, Feb 20, 2003

Genetic information flow



DNA

Gene X

Gene Y

transcription

mRNA

AUG UUU…UAA

translation

Protein Sequence

Met Phe…

Protein Function

Some of the canonical bioinformatics tasks

Centered around the processing of genetic information

Finding genes: exons, introns, transcription start site, splicing signal,…

Inferring the function of protein based on sequence

Understanding gene regulation from sequence signal

Strategy:

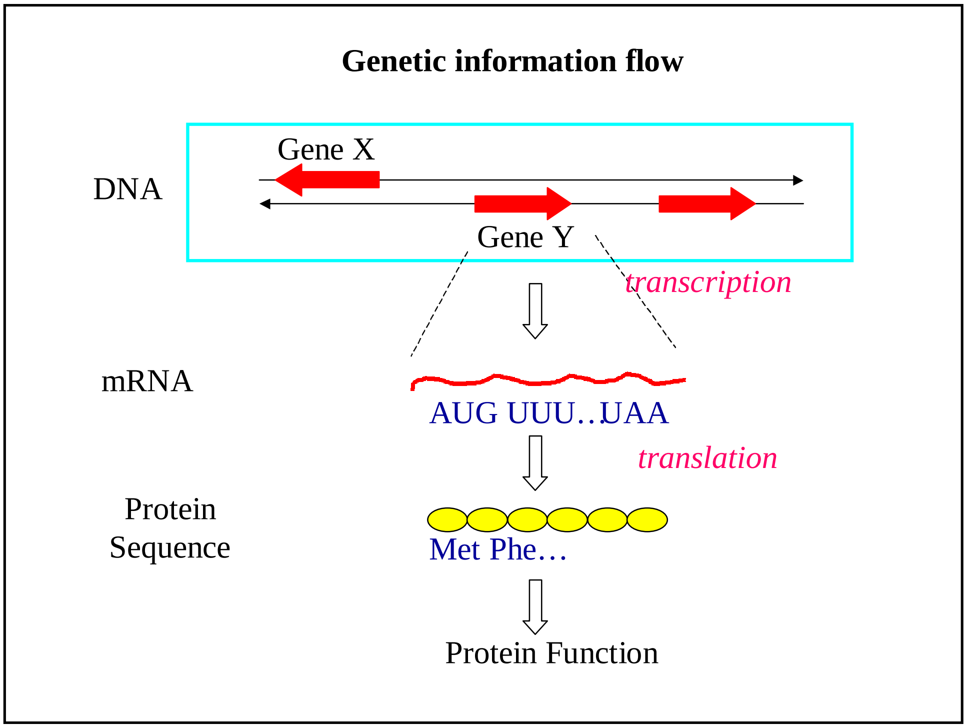
Mimic what the cell does

Use statistics cell can not utilize

Information available

- Primary Sources
  - DNA sequences, GeneBank
  - Protein sequences and structures, SwissProt, PDB
- Genome databases for various organisms
- Other more specialized sources, E.g.,
  - Protein families, domains, structure classification, Pfam, BLOCK,..
  - Promoter databases, EPD
  - Transcription factor binding sites, TRANSFAC

- Large scale experimental data
  - Gene expression (DNA micro-array, proteomics data), SMD…
  - Genomic scale functional assay
  - Protein protein interaction data (e.g., two hybrid screen), DIP…
  - Transcription factor location data (e.g.,CHIP-on-CHIP)



Genetic information flow

Gene X

DNA

Gene Y

transcription

mRNA

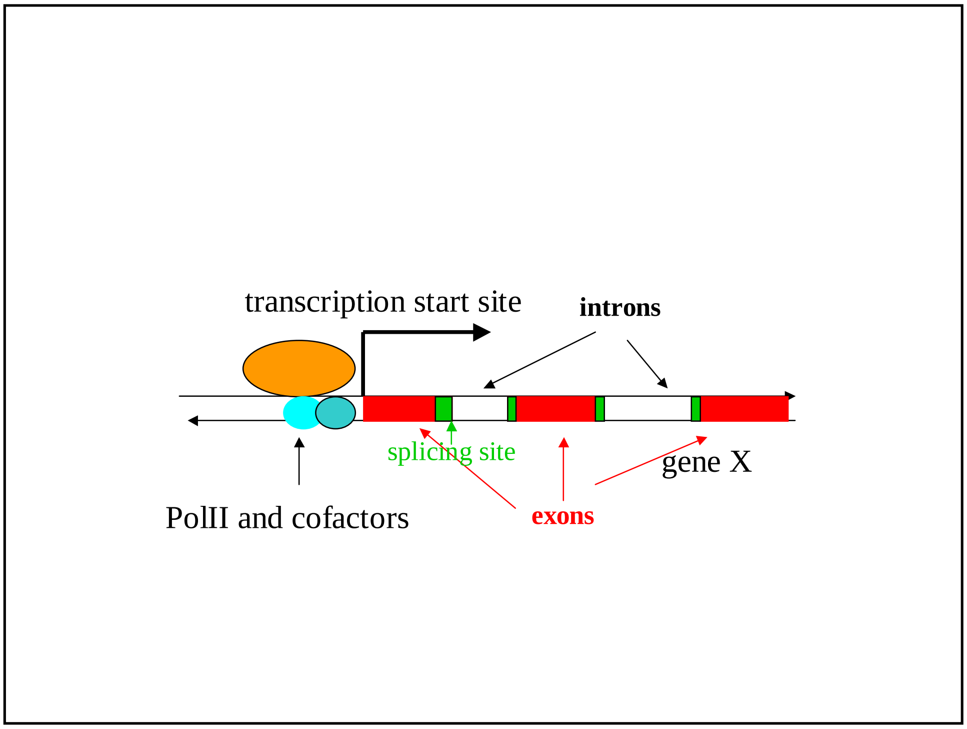
AUG UUU…UAA

translation

Protein Sequence

Met Phe…

Protein Function



transcription start site

introns

splicing site

gene X

PolII and cofactors

exons

Gene finding

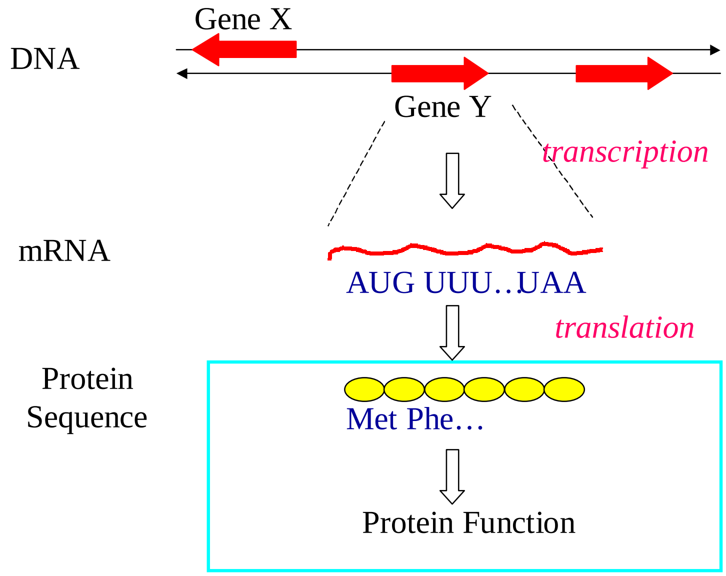
Type of signals

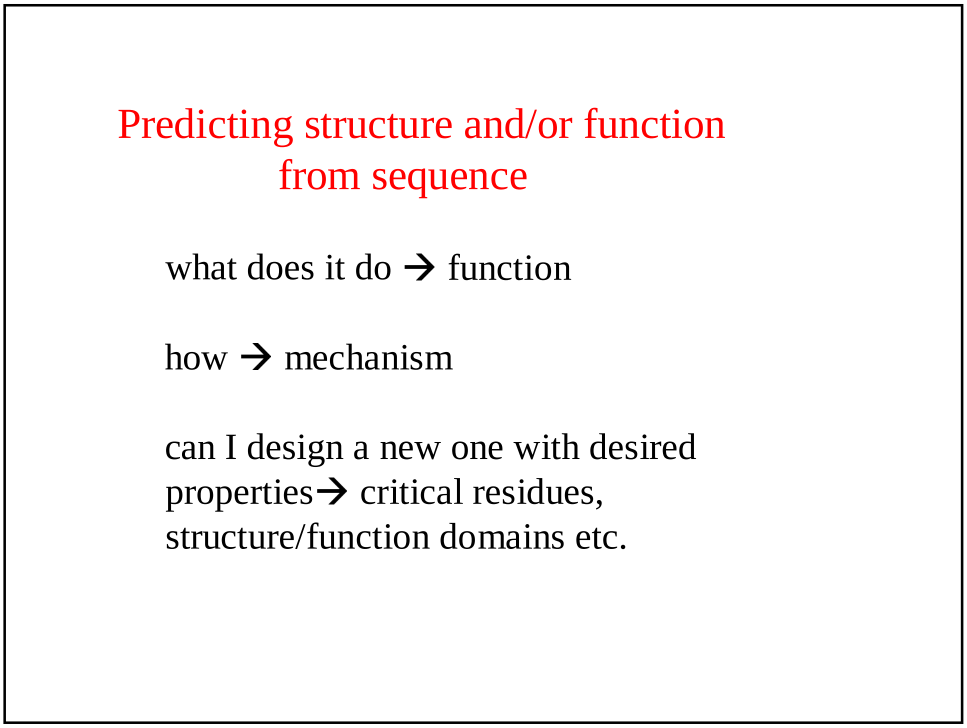
promoter, splicing site, transcription termination, etc
base frequency at diff. codon positions, 6'mer, etc
homology to EST, known protein seq.

Need training set

a list of gene finding algorithms

GENSCAN HMM ~90% single bp, ~80% exon
Grail neural network



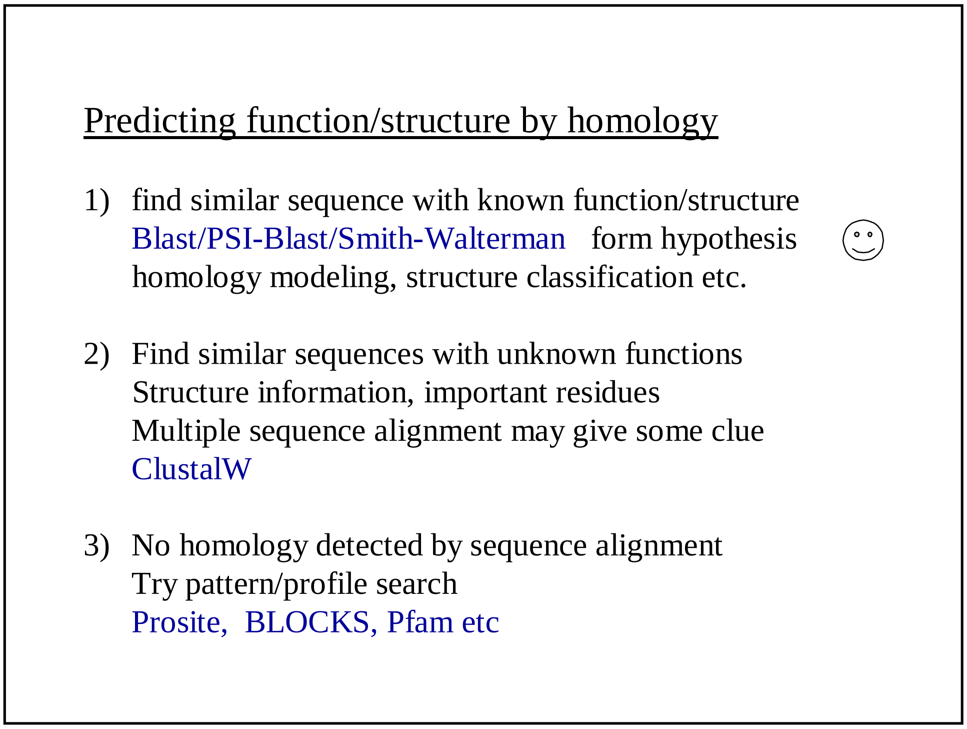


Predicting structure and/or function from sequence

what does it do → function

how → mechanism

can I design a new one with desired properties→ critical residues, structure/function domains etc.



Predicting function/structure by homology

1) find similar sequence with known function/structure
   Blast/PSI-Blast/Smith-Walterman   form hypothesis ☺
   homology modeling, structure classification etc.

2) Find similar sequences with unknown functions
   Structure information, important residues
   Multiple sequence alignment may give some clue
   ClustalW

3) No homology detected by sequence alignment
   Try pattern/profile search
   Prosite, BLOCKS, Pfam etc

sequence alignment

```
DALGKTNAVAHKLLVDD
| |||     ||||  |
DLLGK--VAQHKLLTAD
```

How to score an alignment

match and mismatch: scoring (substitution) matrix

gap penalty: opening a gap; extension of a gap

Significance of an alignment

E value: expected number by chance

Substitution matrices

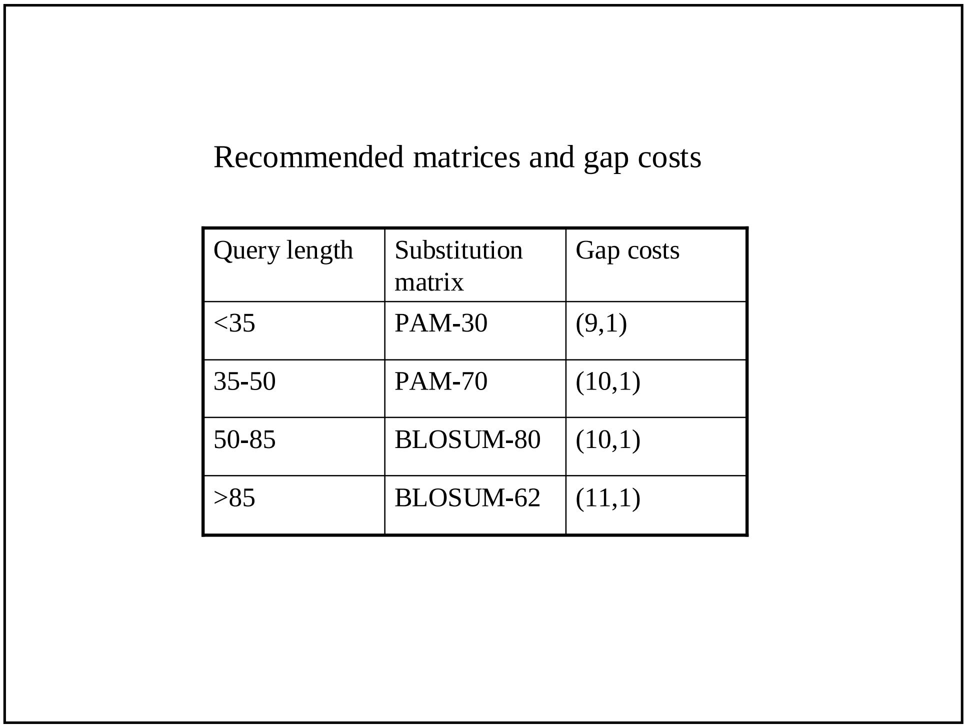
$$s(a,b) = \log\left(\frac{p_{ab}}{q_a q_b}\right)$$

frequency of aligned pairs

frequency of amino acids

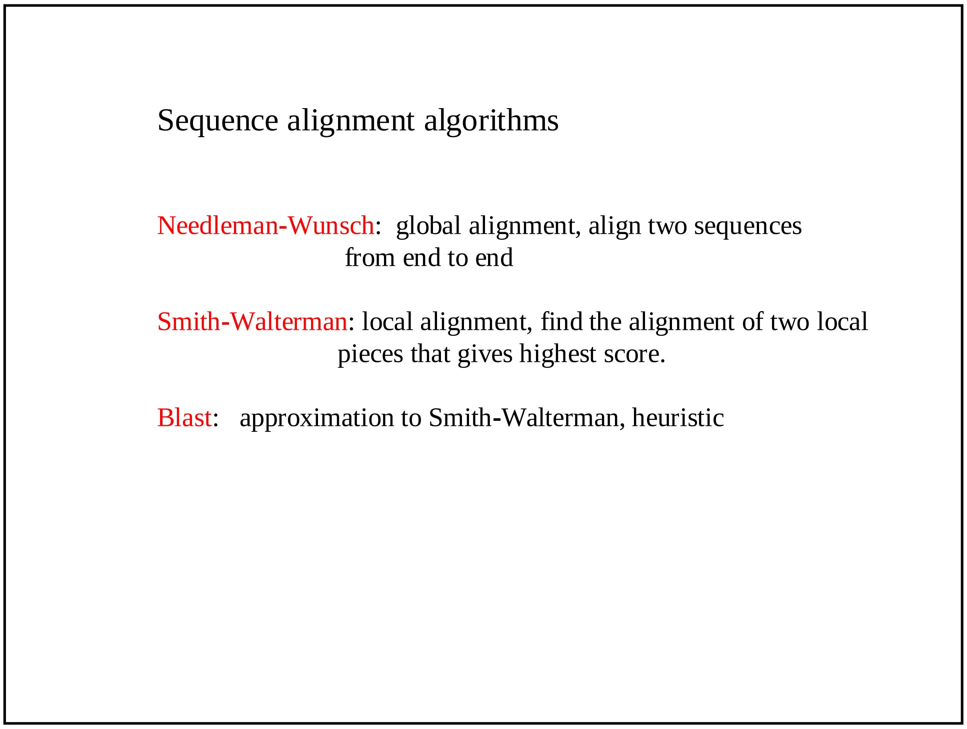
PAM (Dayhoff) matrices: derived from the ungapped alignment of very similar proteins, extrapolate to longer evolutionary distance. PAM-x, larger x corresponds to longer distance

BLOSUM matrices: derived from aligned ungapped regions in the block database. BLOSUM-y, smaller y corresponds to longer distance



Recommended matrices and gap costs

Query length	Substitution matrix	Gap costs
<35	PAM-30	(9,1)
35-50	PAM-70	(10,1)
50-85	BLOSUM-80	(10,1)
>85	BLOSUM-62	(11,1)

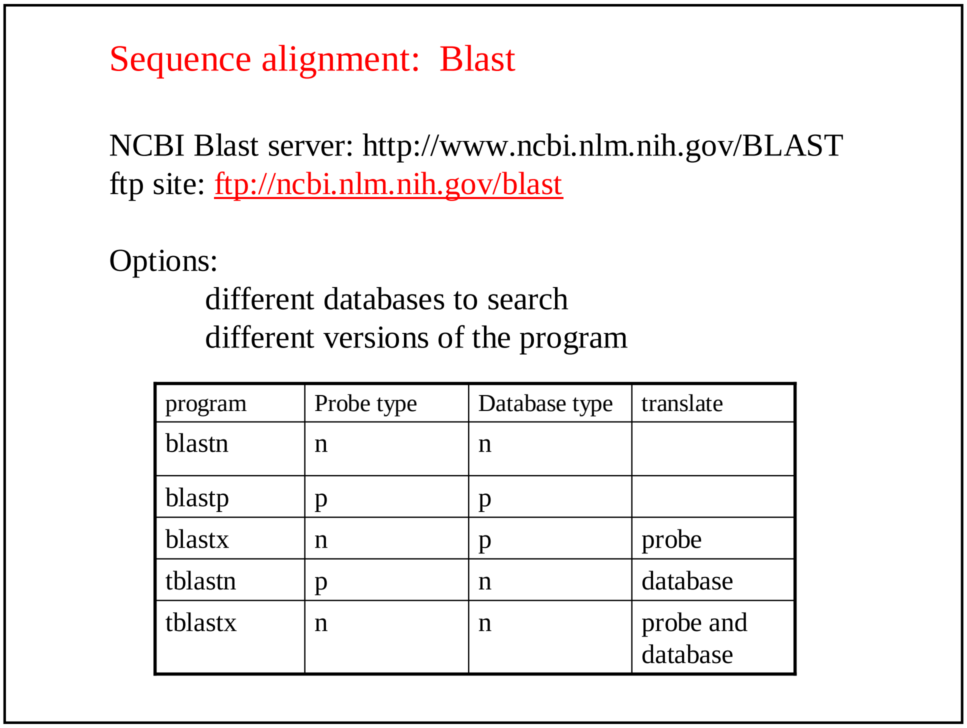


Sequence alignment algorithms

Needleman-Wunsch: global alignment, align two sequences from end to end

Smith-Walterman: local alignment, find the alignment of two local pieces that gives highest score.

Blast: approximation to Smith-Walterman, heuristic



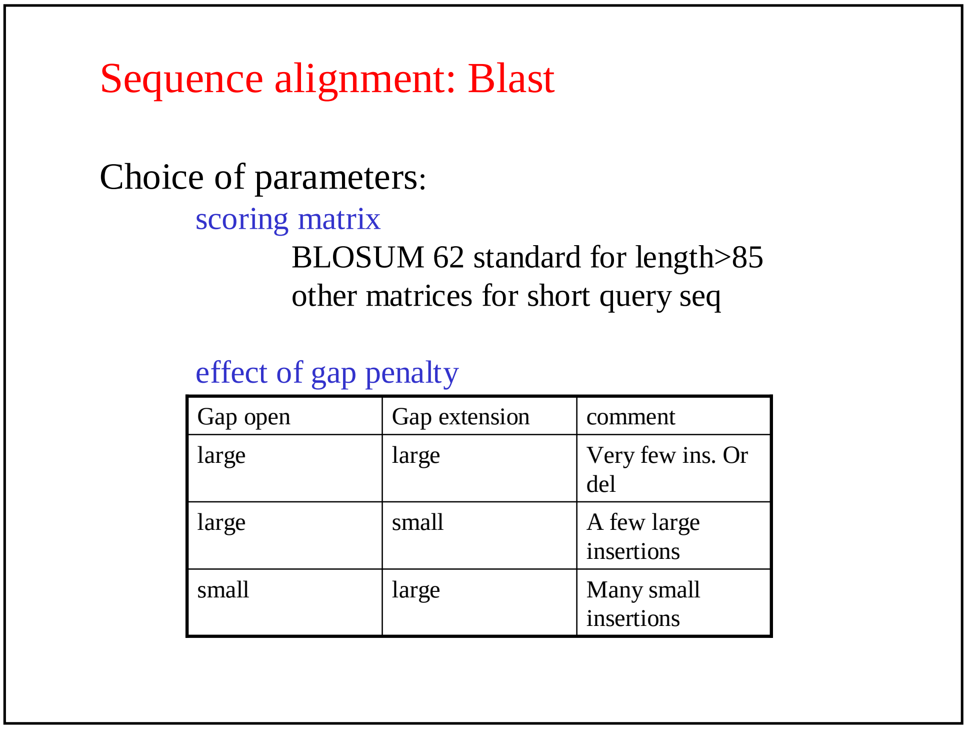
Sequence alignment: Blast

NCBI Blast server: http://www.ncbi.nlm.nih.gov/BLAST
ftp site: ftp://ncbi.nlm.nih.gov/blast

Options:

- different databases to search
- different versions of the program

program	Probe type	Database type	translate
blastn	n	n	
blastp	p	p	
blastx	n	p	probe
tblastn	p	n	database
tblastx	n	n	probe and database



Sequence alignment: Blast

Choice of parameters:

scoring matrix

BLOSUM 62 standard for length>85
other matrices for short query seq

effect of gap penalty

Gap open	Gap extension	comment
large	large	Very few ins. Or del
large	small	A few large insertions
small	large	Many small insertions

Sequence alignment

What is a significant hit

protein: E value < 0.001
length > 80 AA sequence identity > 30%

coding DNA: use blastx, tblastx, translate

noncoding DNA: may use small gap penalty

PSI-Blast

Position Specific Iterated Blast

Using statistically significant hit from Blast
Convert the alignments into position specific score matrix
Search database using the matrix
iterate

Can detect weaker homology

An example
result1 result2

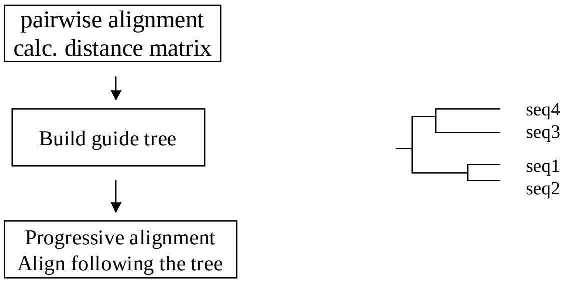
Multiple sequence alignment → structural/evolutionary Relationship (Nature performed mutagenesis for us)

```
sp|P49789|FHIT_HUMAN -------------MSFRFGQHLIKPSVVFLKTELSFALVNRKPVVPGHVLVCPL
sp|O89106|FHIT_MOUSE -------------MSFRFGQHLIKPSVVFLKTELSFALVNRKPVVPGHVLVCPL
sp|P49776|APH1_SCHPO ------------PKQLYFSKFPVG-SQVFYRTKLSAAFVNLKPILPGHVLVIPQ
sp|P49775|HNT2_YEAST MILSKTKKPKSMNKPIYFSKFLVT-EQVFYKSKYTYALVNLKPIVPGHVLIVPL
sp|Q11066|YHIT_MYCTU -----------MPCVFCAIIAGEAPAIRIYEDGGYLAILDIRPFTRGHTLVLPK
sp|Q58276|Y866_METJA ------------MCIFCKIINGEIPAKVVYEDEHVLAFLDINPRNKGHTLVVPK
```

Multiple sequence alignment

Generalization of dynamical programming works only for small number (<8) of short sequences

Progressive alignment: practical algorithms for a large number of sequences



pairwise alignment
calc. distance matrix

Build guide tree

Progressive alignment
Align following the tree

Multiple sequence alignment

clustalW

sequence weighting according to divergence
substitution matrix chosen based on expected similarity
carefully fine tuned gap penalties

position specific, residue specific

example

result

Searching local profiles

using profile databases

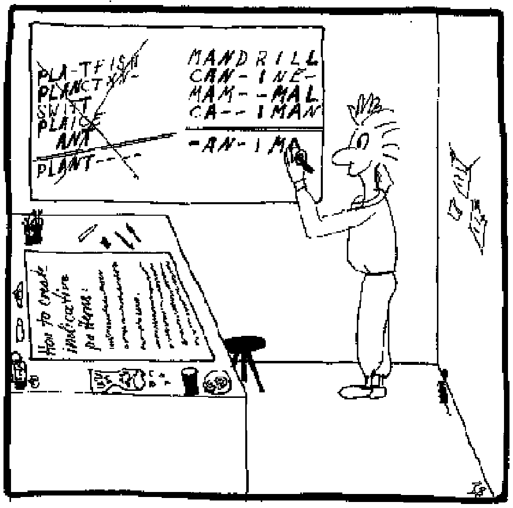
Prosite: patterns and profiles from conserved motifs, functional sites
1580 patterns and profiles (release 17.28 Nov 2002)

BLOCKS: based on automatic ungapped alignment, position specific scoring matrices
2101 groups 8656 blocks (version 13.0)

Pfam: multiple sequence alignment and profile HMMs
models for 4832 protein families (Oct, 2002)

Several others

How we develop Prosite patterns!



Brigitte Boeckmann / 1995

An example of pattern/profile databases

Prosite:

originally only regular expression patterns
now include profiles
e.g., build pattern around an active residue

```
ALRDFATHDDF
SMTAEATHDSI
ECDQAATHEAS

CONSENSUS PATTERN ATH[D,E]
```

BLOCKS:

Construct position specific scoring matrix (PSSM)
based on gapless blocks of multiply aligned sequences

$$w(x,a) = \sum_{b} f(x,b) \times s(b,a)$$

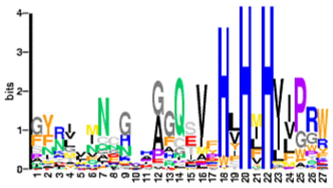
f(x, b) weighted frequency of b at position x
s(b, a) is the scoring matrix. there are more elaborated
Method based on position dependent pseudo-counts

For example,

```
….VGAHA
….VNVDE
….VEADE
….FNANP
….IAGAD
```

M(1, a) = 3/5*s(V,a) + 1/5*s(F,a) + 1/5*s(I,a)

Example of blocks profile
Histidine triad
Zinc binding motif



PSSM of IPB001310B (HIT;) 91 sequences.

Predicting structure from sequence

transmembrane helices 80%~95% accuracy

windows of 20 hydrophobic AA, database of known helices

http://www.ch.embnet.org/software/TMPRED_form.html

secondary structure prediction 70% ~80%

tertiary structure prediction ?

homology modeling, fold recognition, ab initio

homology modeling:

http://www.expasy.ch/swissmod/SWISS-MODEL.html

Secondary structure prediction

single sequence based

Chou and Fasman helix propencity

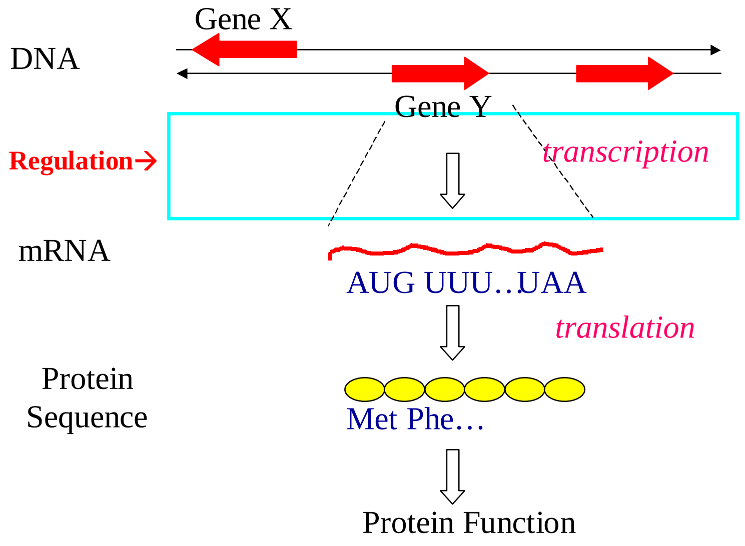
nnpred (Kneller, Cohen, Langridge) neural nets based

….

Algorithms using multiple sequence information

PHD

….



Gene regulation

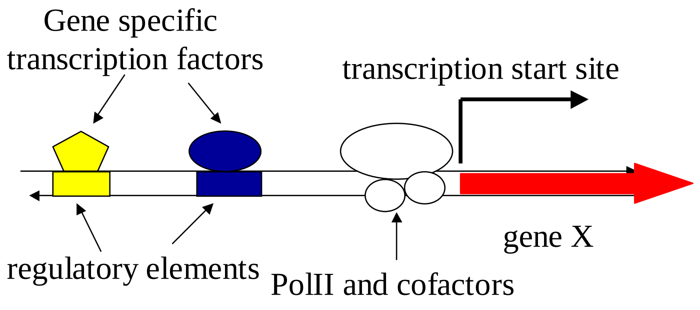
The expression of gene can be controlled in various ways

transcriptional: the amount of mRNA synthesized
translational: the amount of proteins made
post-translational: modify proteins

Transcriptional control is a major mechanism for regulation

transcriptional program is encoded in genomic DNA

Transcriptional regulation



Gene specific transcription factors

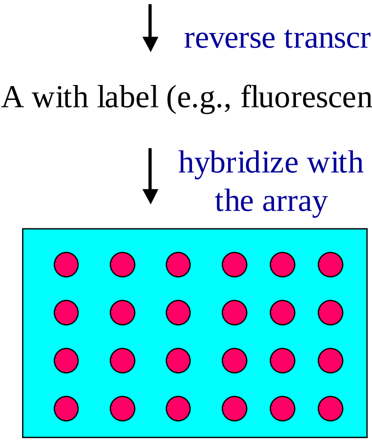
transcription start site

gene X

regulatory elements

PolII and cofactors

Measure genome-wide gene expression



mRNA from cell extract

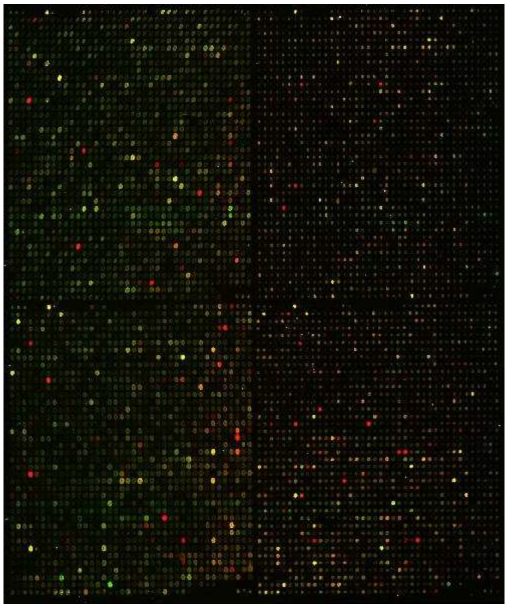
reverse transcribe

cDNA with label (e.g., fluorescence)

hybridize with
the array

DNA microarray

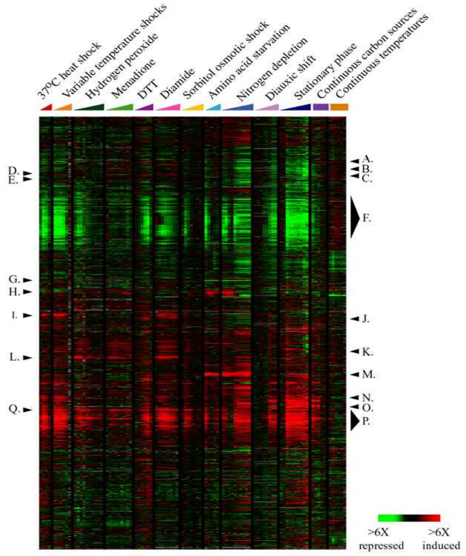
DNA probes for
all the genes



Derisi et al
Diauxic shift

gene regulation → cellular response to environment

Figure 1.



Gasch et al.

complete genome sequences

+

genome-wide expression data
(e.g., from DNA microarray)

↓

opportunity to decipher a cell’s
regulatory program

General approaches

Knowledge based approach

collect examples of known TF binding sites → databases
build profiles of TF recognition sites → search query seq.

Discover novel regulatory sites

pattern recognition

finding patterns in a group of co-regulated genes
identifying combinatorial motifs from the whole genome

knowledge based: Databases of TF binding sites

TRANSFAC: The Transcription Factor Database
a collection of experimentally determined
transcription factor binding sites
Release IV
site: 8415
gene: 1078
factor: 2785
profiles: 309 matrices

SCPD: TF binding sites database for Yeast
~ 500 sites representing ~100 factors

E Coli: Church lab website
~ 800 sites representing ~ 60 factors

Knowledge based: building TF recognition profiles

Align examples of binding sites for a given TF (MCB)

```
>YDL102W   ACGCGT
>YDL164C   ACGCGT
>YDL164C   ACGCGA
>YJL194W   ACGCGT
>YJL194W   ACGCGA
>YMR199W   CCGCGT
>YMR199W   TCGCGA
>YMR199W   ACGCGT
>YNL102W   ACGCGT
>YNL102W   ACGCGT
>YOR074C   ACGCGT
>YOR074C   ACGCGT
```

A	**10**	**0**	**0**	**0**	**0**	**3**
T	**1**	**0**	**0**	**0**	**0**	**9**
G	**0**	**0**	**12**	**0**	**12**	**0**
C	**1**	**12**	**0**	**12**	**0**	**0**

← Alignment matrix

position specific probability matrix

	1	2	3	4	5	6
A	0.84	0.00	0.00	0.00	0.00	0.25
C	0.08	1.00	0.00	1.00	0.00	0.00
G	0.00	0.00	1.00	0.00	1.00	0.00
T	0.08	0.00	0.00	0.00	0.00	0.75

⇐ $f_{i,\sigma}$

probability of certain base occurring in the binding site is given by the above matrix

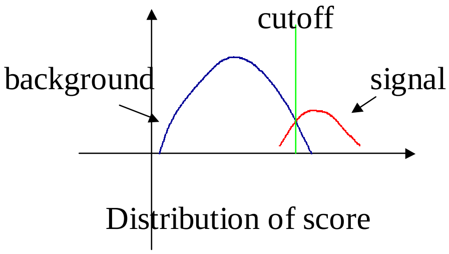
probability of observing certain base not in the binding site is given by the background frequency f_{σ}^{0}

Predicting new sites using the matrix

$$w_{i\sigma} = \ln[f_{i\sigma} / f_{\sigma}^{0}]$$

Score a sequence using the position weight matrix

AGACGT → $w_{1A} + w_{2C} + w_{3A} + w_{4T} + w_{5G} + w_{6A}$ Log likelihood ratio



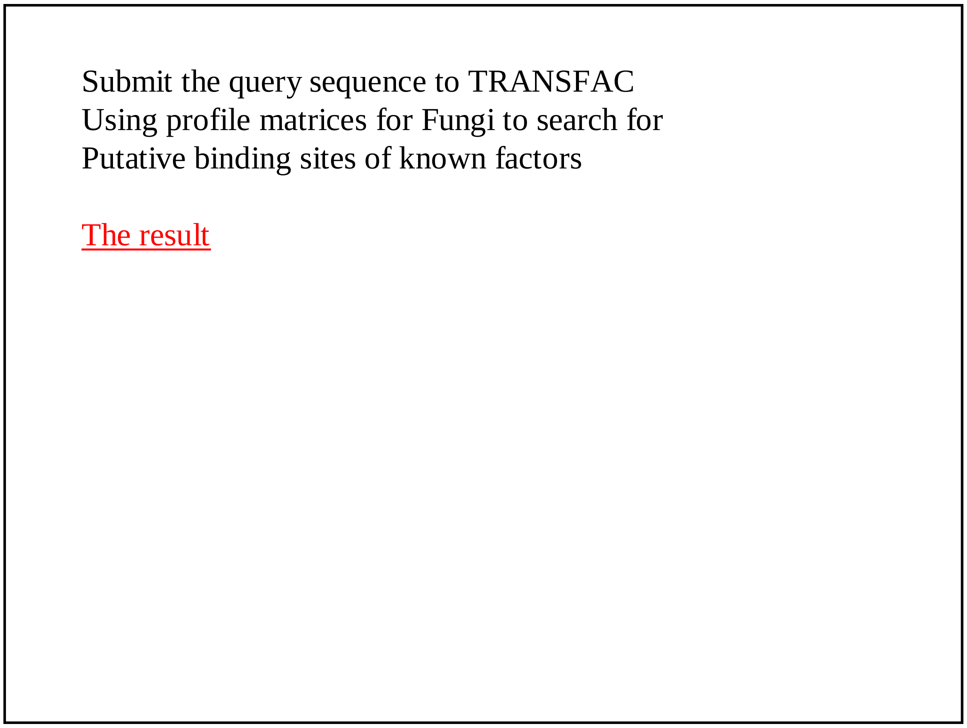
Distribution of score

Knowledge based: searching for known sites in query seq.

Example: his4 histidine biosynthesis

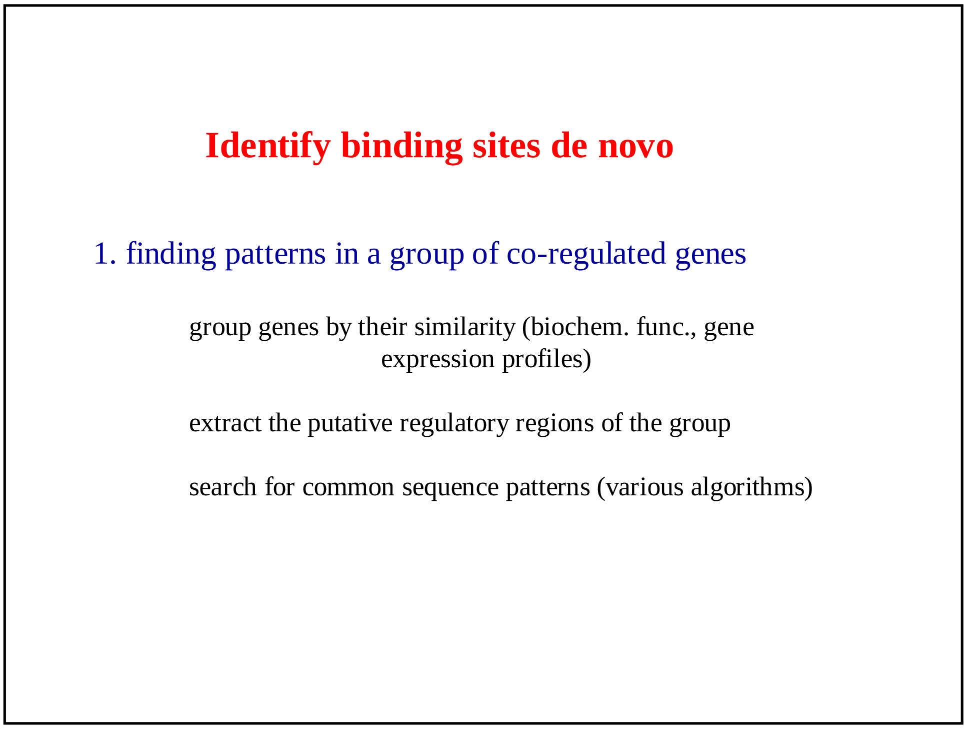
Extract the putative regulatory region for his4, 600bp 5’ of the coding
region

```
gggctaaagaacgcgaacaattgaaaatgcataacgattcgctcagtaaagaatacc
aaaatttgagcaaggaactatttttgacaaaaccacaagattcctcatcggaagaggtg
gcatccttaacgaaaaaacttgaagaggctaatgaaaaaatcaaacagttggaacagg
ctcaagcacaaacagccgtggaatcgttgccaattttcgacccccctgcaccagtcgata
ccacggcaggaagacaacagtggtgtgagcattgcgatacgatgggtcataatacagca
gaatgcccccatcacaatcctgacaaccagcagttcttctaggcagtcgaactgactctaat
agtgactccggtaaattagttaattaattgctaaacccatgcacagtgactcacgtttttttatca
gtcattcgatatagaaggtaagaaaaggatatgactatgaacagtagtatactgtgtatataat
agatatggaacgttatattcacctccgatgtgtgttgtacatacataaaaatatcatagcacaa
ctgcgctgtgtaatagtaatacaatagtttacaaaattttttttctgaata
```



Submit the query sequence to TRANSFAC
Using profile matrices for Fungi to search for
Putative binding sites of known factors

The result



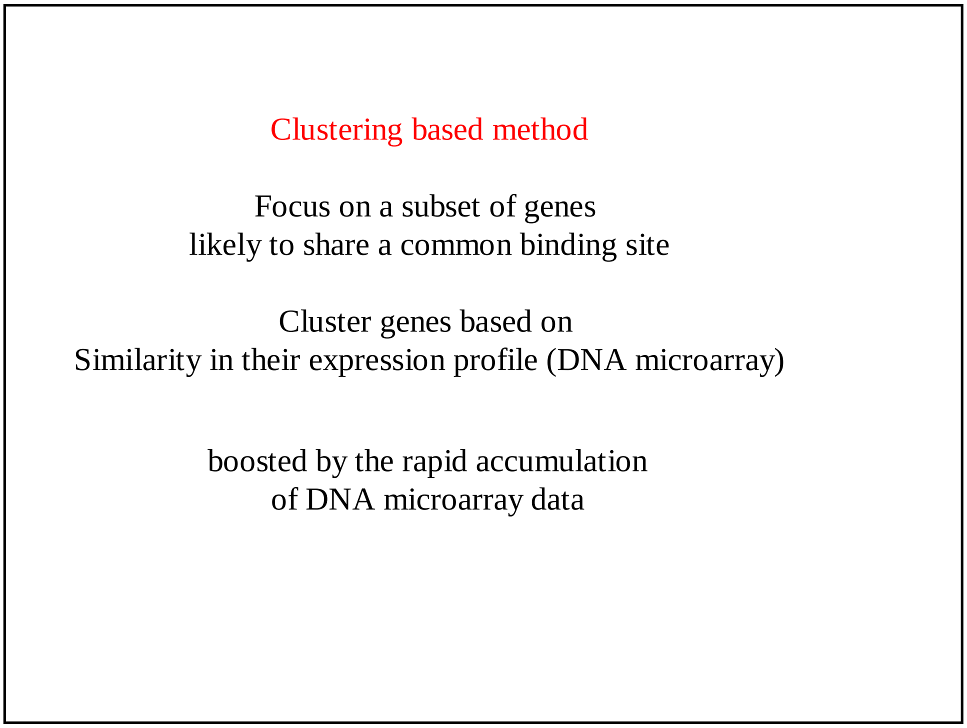
Identify binding sites de novo

1. finding patterns in a group of co-regulated genes

group genes by their similarity (biochem. func., gene expression profiles)

extract the putative regulatory regions of the group

search for common sequence patterns (various algorithms)

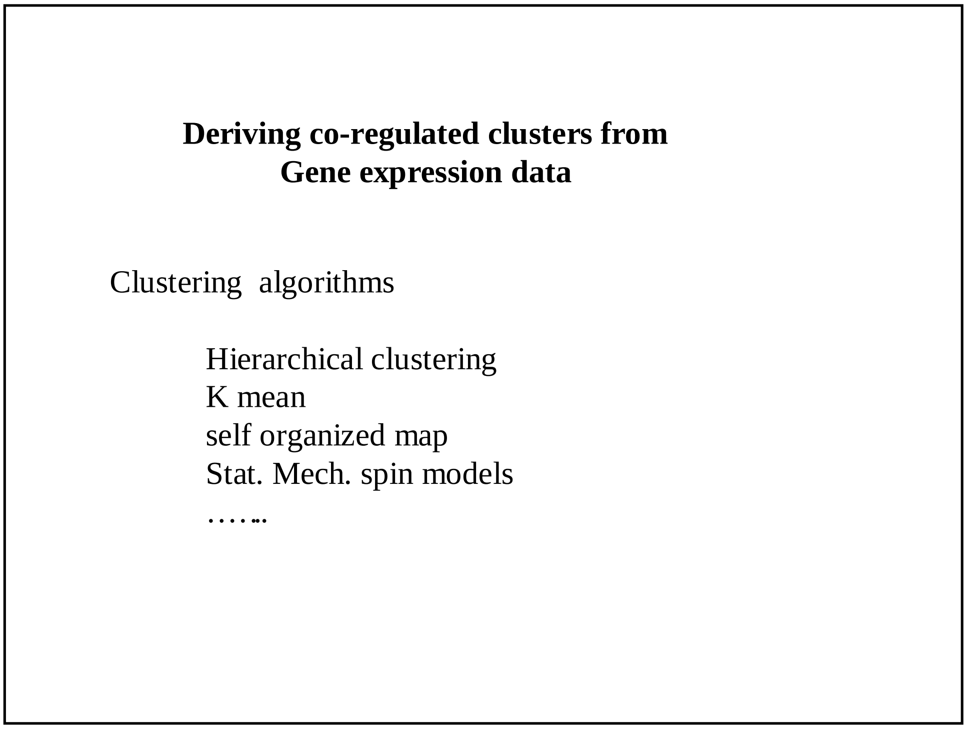


Clustering based method

Focus on a subset of genes
likely to share a common binding site

Cluster genes based on
Similarity in their expression profile (DNA microarray)

boosted by the rapid accumulation
of DNA microarray data



**Deriving co-regulated clusters from
Gene expression data**

Clustering algorithms

Hierarchical clustering
K mean
self organized map
Stat. Mech. spin models
……..

Hierarchical clustering (Eisen et. al.)

Define similarity between a pair of genes X and Y

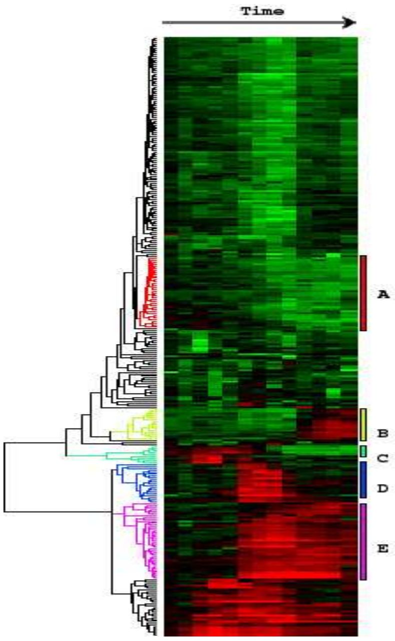
Xi: expression level of X in experiment i
Yi: expression level of X in experiment i

$$s(X,Y) = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{N\sigma_X \sigma_Y}$$

Assemble genes into a dendrogram based on the pairwise similarities

Example from Eisen et al.

human fibroblasts
serum stimulated



Subsets of genes defined by the clusters

detecting common sequence patterns in
the regulatory sequences of the subset of genes

❑ frequency count of substrings or regular expression patterns

❑ local multiple sequence alignment

Algorithm based on frequency contrast

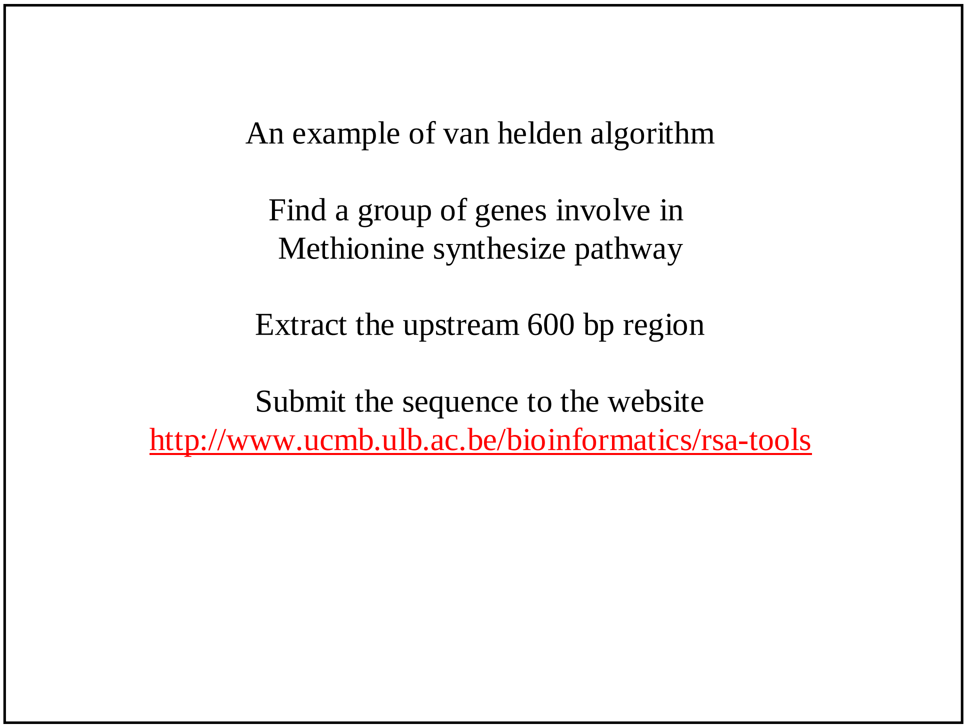
oligonucleotides frequency contrast

(van Helden, Andre, and Collado-vides)

count the number in the dataset

count the number in a contrast dataset (e.g., whole genome promoters)

use the expected frequency from contrast set to calculate significance of over-representation

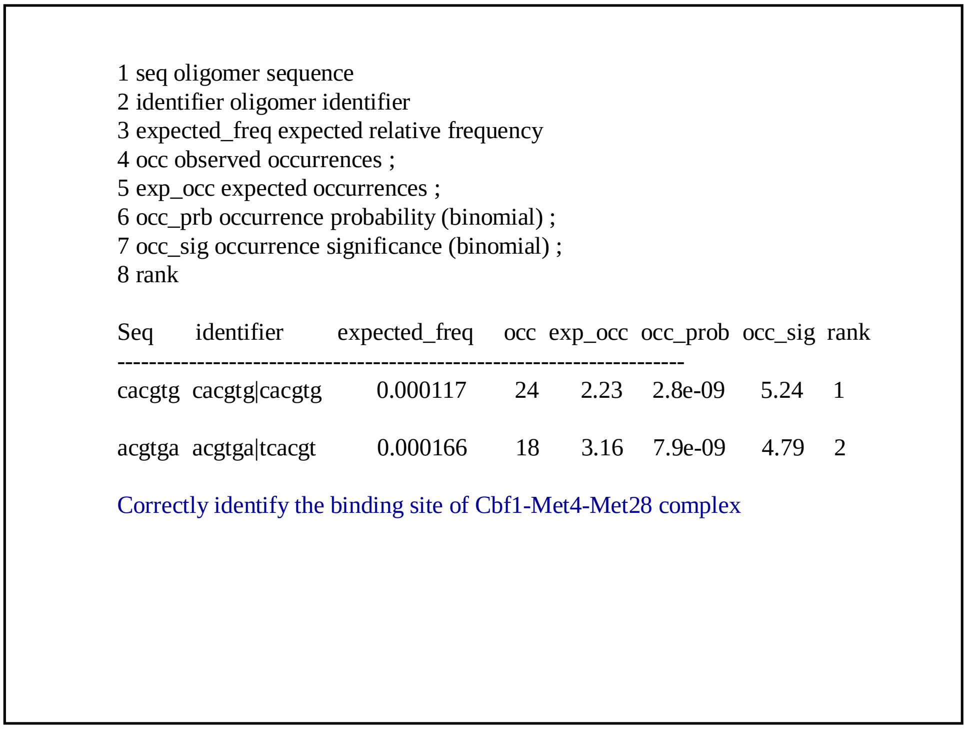


An example of van helden algorithm

Find a group of genes involve in
Methionine synthesize pathway

Extract the upstream 600 bp region

Submit the sequence to the website
http://www.ucmb.ulb.ac.be/bioinformatics/rsa-tools



1 seq oligomer sequence
2 identifier oligomer identifier
3 expected_freq expected relative frequency
4 occ observed occurrences ;
5 exp_occ expected occurrences ;
6 occ_prb occurrence probability (binomial) ;
7 occ_sig occurrence significance (binomial) ;
8 rank

```
Seq     identifier      expected_freq   occ  exp_occ  occ_prob  occ_sig  rank
---------------------------------------------------------------------------
cacgtg  cacgtg|cacgtg       0.000117     24     2.23   2.8e-09     5.24    1

acgtga  acgtga|tcacgt       0.000166     18     3.16   7.9e-09     4.79    2
```

Correctly identify the binding site of Cbf1-Met4-Met28 complex

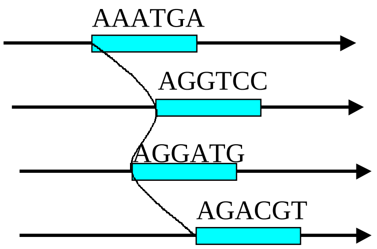
Algorithms based on multiple local alignments

Consensus (Hertz and Stormo)

Gibbs-Sampler (Lawrence, Liu, Newald et al.)

MEME (Baily and Elkan)

Solving the problem
A model for the motif



alignment matrix

	1	2	3	4	5	6
A	4	1	2	1	0	1
C	0	0	0	1	1	1
G	0	3	2	0	2	1
T	0	0	0	2	1	1

position specific frequency matrix

$$\begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline A & 1.00 & 0.25 & 0.50 & 0.25 & 0.00 & 0.25 \\ C & 0.00 & 0.00 & 0.00 & 0.25 & 0.25 & 0.25 \\ G & 0.00 & 0.75 & 0.50 & 0.00 & 0.50 & 0.25 \\ T & 0.00 & 0.00 & 0.00 & 0.50 & 0.25 & 0.25 \end{array} \quad \Leftarrow \quad f_{i\sigma}$$

↑

Information 2 bits

$$Information = \sum_{i,\sigma} f_{i\sigma} \ln[\, f_{i\sigma} \,/\, f^{0}_{\sigma} \,]$$

Alignment based algorithms
MEME, Consensus, Gibbs Sampler
Search for alignment path → optimize information

Statistical Model:

probability of observing certain base inside
the motif is given by the position-specific prob. matrix

probability of observing certain base outside
the motif is given by the background frequency

Maximum likelihood formulation

Starting positions of the motif unknown $\vec{x} = (x_1, x_2, ..., x_N)$

Position specific probability matrix unknown $f_{i,\sigma}$

need to be inferred from the observed sequence data

$$P(seq \mid f_{i,\sigma}, \vec{x}) = \prod_{i=1}^{N} \left(\prod_{j=1}^{x_i - 1} f^0_{\sigma_{ij}} \prod_{j=x_i}^{x_i+w-1} f_{j-x_i+1,\sigma_{ij}} \prod_{j=x_i+w}^{L} f^0_{\sigma_{ij}} \right)$$

N Number of sequences

L Length of the sequence

w Width of the motif

σ_{ij} Base of sequence i at position j

$$P(seq \mid f_{i,\sigma}, \vec{x}) = const \prod_{j=1}^{w} \prod_{\sigma} \left(\frac{f_{j,\sigma}}{f^0_{\sigma}} \right)^{n_{j,\sigma}(\vec{x})}$$

likelihood ratio

$n_{j,\sigma}(\vec{x})$ Total number of count for base σ at position j in the alignment

Maximizing $P(seq, \vec{x} \mid f_{i,\sigma})$ w.r.t. $f_{i,\sigma}$ With $\vec{x}$ fixed

$$\log P(seq, \vec{x} \mid \hat{f}_{i,\sigma}) = N \sum_{j=1}^{w} \hat{f}_{j,\sigma} \log \left(\frac{\hat{f}_{j,\sigma}}{f^0_{\sigma}} \right)$$

log likelihood ratio
relative entropy

$$\hat{f}_{j,\sigma} = \frac{n_{j,\sigma}(\vec{x})}{\sum_{\sigma} n_{j,\sigma}(\vec{x})}$$

in reality, this formula is modified by adding pseudo counts due to Baysian estimate

Then maximize the above relative entropy w.r.t $\vec{x}$

→ Alignment path.

Stormo-Hartzell Algorithm: Consensus

- each of the length w substrings of the first sequence are aligned against all the substrings of the same length in the second sequence, matrices derived, N top matrices with highest information contents are saved
- the next sequence on the list is added to the analysis, all the matrices saved previously are paired with the substrings of the added sequence and top N matrices saved
- repeat the previous step until all the sequences have been processed

Example: 22 genes identified as pho4 target by microarray, O’shea lab

```
YAR071W:600:-600
\catcaagatgagaaaataaagggatttttcgttcttttatcattttctctttctcacttccgactacttcttatatctactttcatcgtttcattcatcgtgggtgtctaataaagtttta
atgacagagataaccttgataagctttttcttatacgctgtgtcacgtatttattaaattaccacgttttcgcataacattctgtagttcatgtgtactaaaaaaaaaaaaaaaaa
gaaataggaaggaaagagtaaaaagttaatagaaaacagaacacatccctaaacgaagccgcacaatcttggcgttcacacgtgggtttaaaaaggcaaattacacag
aatttcagaccctgtttaccggagagattccatattccgcacgtcacattgccaaattggtcatctcaccagatatgttatacccgttttggaatgagcataaacagcgtcgaa
ttgccaagtaaaacgtatataagctcttacatttcgatagattcaagctcagtttcgccttggttgtaaagtaggaagaagaagaagaagaagaggaacaacaacagcaaa
gagagcaagaacatcatcagaaatacca\
YBR092C:600:-600
\aatcaatgacttctacgactatgctgaaaagagagtagccggtactgacttcctaaaggtctgtaacgtcagcagcgtcagtaactctactgaattgaccttctactgggac
tggaacactactcattacaacgccagtctattgagacaatagttttgtataactaaataatattggaaactaaatacgaatacccaaatttttatctaaattttgccgaaagatta
aaatctgcagagatatccgaaacaggtaaatggatgtttcaatccctgtagtcagtcaggaacccatattatattacagtattagtcgccgcttaggcacgcctttaattagca
aaatcaaaccttaagtgcatatgccgtataagggaaactcaaagaactggcatcgcaaaaatgaaaaaaaggaagagtgaaaaaaaaaaaattcaaaagaaatttacta
aataataccagtttgggaaatagtaaacagctttgagtagtcctatgcaacatatataagtgcttaaatttgctggatggaagtcaattatgccttgattatcataaaaaaaata
ctacagtaaagaaagggccattccaaattacct\
YBR093C:600:-600
\cgctaatagcggcgtgtcgcacgctctctttacaggacgccggagaccggcattacaaggatccgaaagttgtattcaacaagaatgcgcaaatatgtcaacgtatttgg
aagtcatcttatgtgcgctgctttaatgttttctcatgtaagcggacgtcgtctataaacttcaaacgaaggtaaaaggttcatagcgctttttctttgtctgcacaaagaaatata
tattaaattagcacgttttcgcatagaacgcaactgcacaatgccaaaaaaagtaaaagtgattaaaagagttaattgaataggcaatctctaaatgaatcgatacaaccttg
gcactcacacgtgggactagcacagactaaatttatgattctggtccctgttttcgaagagatcgcacatgccaaattatcaaattggtcaccttacttggcaaggcatatac
ccatttgggataagggtaaacatctttgaattgtcgaaatgaaacgtatataagcgctgatgttttgctaagtcgaggttagtatggcttcatctctcatgagaataagaacaa
caacaaatagagcaagcaaattcgagattacca\
YBR296C:600:-600
\gaaatctcggtttcacccgcaaaaaagtttaaatttcacagatcgcgccacaccgatcacaaaacggcttcaccacaagggtgtgtggctgtgcgatagacctttttttctt
tttctgctttttcgtcatccccacgttgtgccattaatttgttagtgggcccttaaatgtcgaaatattgctaaaaattggcccgagtcattgaaaggctttaagaatataccgtac
aaaggagtttatgtaatcttaataaattgcatatgacaatgcagcacgtgggagacaaatagtaataatactaatctatcaatactagatgtcacagccactttggatccttcta
ttatgtaaatcattagattaactcagtcaatagcagatttttttacaatgtctactgggtggacatctccaaacaattcatgtcactaagcccggttttcgatatgaagaaaattat
atataaacctgctgaagatgatctttacattgaggttattttacatgaattgtcatagaatgagtgacatagatcaaaggtgagaatactggagcgtatctaatcgaatcaatat
aaacaaagattaagcaaaaatg\
```

Consensus output for Pho4 regulated genes

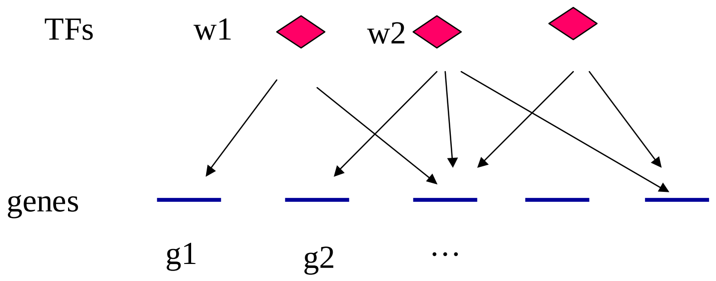
```
MATRIX 1
number of sequences = 22
information = 8.80903
ln(p-value) = -153.757   p-value = 1.67566E-67
ln(expected frequency) = -13.357   expected frequency = 1.58165E-06

A |  6   5 20  3   0  3  0   0  0   6
G | 11  0   0  5 22  0 21 15 14  2
C |  4 17   0 14  0  0   1  2  8   1
T |  1  0   2  0  0 19  0  5  0 13
     G  C   A  C  G  T  G  G  G  T

  1|1   :  1/317   ACACGTGGGT
  2|2   :  2/55    AAAGGTCTGT
  3|3   :  3/347   ACACGTGGGA
  4|4   :  4/274   GCACGTGGGA
  5|5   :  5/392   CAACGTGTCT
  6|6   :  6/395   ACAAGTGGGT
  7|7   :  7/321   ACACGTGGGA
  8|8   :  8/536   GCAAGTGGCT
  9|9   :  9/177   GCTGGTGTGT
 10|10  :  10/443   GCACGTGTCT
 11|11  :  11/14    CCAGGTGCCT
 12|12  :  12/502   GAAAGAGGCA
 13|13  :  13/354   GCACGAGGGA
 14|14  :  14/257   GCACGTGCGA
 15|15  :  15/358   TCACGTGTGT
 16|16  :  16/316   ACACGTGGGT
 17|17  :  17/479   GCACGTGGCT
 18|18  :  18/227   GATGGTGGCT
 19|19  :  19/186   GCACGTGGGG
 20|20  :  20/326   GAAGGAGGGG
 21|21  :  21/307   CCACGTGGGC
 22|22  :  22/255   CCACGTGGCT
```

2. Identifying combinatorial motifs

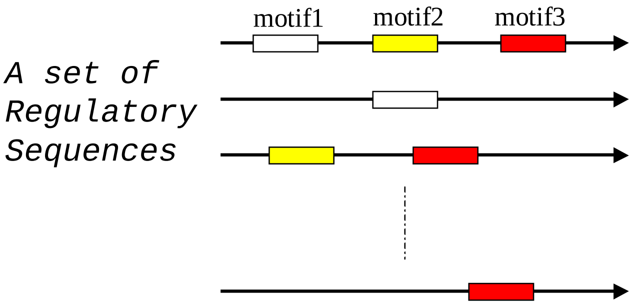
dictionary approach: finding words from a scrambled text

Mobydick, http://genome.ucsf.edu:8080/mobydick

Cell’s regulatory network is complicated



Combinatorial control



How do we find these motifs?

chapterptgpbqdrftezptqtasctmvivwpecjsnisrmbtqlmlfvetl
loomingsfkicallxjgkmekysjerishmaeljplfsomeylqyearstvh
njbagoaxhjtjcokhvneverpmqpmindhowzrbdlzjllonggbhqi
preciselysunpvskepfdjktcgarwtnxybgcvdjfbnohavinglittl
ezorunozsoyapmoneyyvugsgtsqintmyteixpurseiwfmjwgj
nyyveqxwftlamnbxkrsbkyandrnothingcgparticularwtzao
qsjtnmtoqsnwvxfiupinterestztimebymonlnshoreggditho
ughtyxfxmhqixceojjzdhwouldsailpcaboutudxsbsnewtpg
gvjaasxmsvlittleplvcydaowgwlbzizjlnzyxandzolwcudthjd
osbopxkkfdosxardgcseebbthefzrsskdhmawateryjikzicim
ypartmofprtheluworldvtoamfutitazpisagwewayrqbkiosh
avebojwphiixofprmalungipjdrivingpkuyoikrwxoffodhicb
nimtheixyucpdzacemspleenqbpcrmhwvddyaiwnandada
bkpgzmptoregulatingeetheslcirculationvsuctzwvfyxstuzr
dfwvgygzoejdfmbqescwheneverpitfindfmyselfcgrowingne
ostumrydrrthmjsmgrimcczhjmgbkwczoaboutjbwanbwzq
thehrjvdrcjjgmouthuutwheneveritddfouishlawwphxnae

Moby Dick: CHAPTER 1
Loomings

Call me Ishmael. Some years ago- never mind how long precisely-having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen and regulating the circulation. Whenever I find myself growing grim about the mouth; whenever it is a damp, drizzly November in my soul; whenever I find myself involuntarily pausing before coffin warehouses, and bringing up the rear of every funeral I meet; and especially whenever my hypos get such an upper hand of me, that it requires a strong moral principle to prevent me from deliberately stepping into the street, and methodically knocking people' s hats off-then, I account it high time to get to sea as soon as I can.

the Model:

Probabilistic Segmentation/Maximum likelihood

A probabilistic dictionary
Words
probabilities

$$\left\{\begin{array}{lll} \mathrm{A} & \rightarrow & P_{A} \\ \mathrm{C} & \rightarrow & P_{C} \\ \mathrm{G} & \rightarrow & P_{G} \\ \mathrm{T} & \rightarrow & P_{T} \\ \mathrm{GC} & \rightarrow & P_{GC} \\ \mathrm{TATAA} & \rightarrow & P_{TATAA} \end{array}\right\}$$

A | G | T | A | T | A | A | G | C
A | G | T A T A A | G C
A | G | T A T A A | G | C

maximizing
the likelihood
function

$$Z = \sum_{Seg} P_{w_1} P_{w_2} P_{w_3} \ldots P_{w_n}$$

```
Ditionary1                 Ditionary2                 Dictionary3
----------------------------------------------------------------------
e       0.065239           e       0.048730           e       0.042774
t       0.055658           s       0.042589           s       0.040843
a       0.052555           a       0.040539           a       0.038595
o       0.050341           t       0.040442           i       0.036897
n       0.049266           i       0.038550           t       0.036871
i       0.048101           d       0.038547           d       0.036323
s       0.047616           o       0.036486           l       0.035336
h       0.047166           l       0.036300           c       0.034818
r       0.043287           g       0.034509           m       0.034650
l       0.041274           r       0.034496           y       0.034482
d       0.039461           c       0.033916           b       0.034396
u       0.034742           m       0.033724           r       0.034105
m       0.034349           n       0.033321           p       0.034044
g       0.034001           y       0.033227           w       0.033819
w       0.033967           p       0.033156           n       0.033817
c       0.032934           f       0.032863           g       0.033676
f       0.032597           b       0.032780           f       0.033534
y       0.031776           w       0.032009           o       0.033206
p       0.031711           h       0.031494           h       0.033200
b       0.031409           v       0.030727           k       0.032103
v       0.028268           k       0.030445           v       0.031498
k       0.028113           u       0.030379           j       0.031209
j       0.026712           j       0.029268           u       0.031186
q       0.026561           z       0.028905           z       0.031003
z       0.026542           x       0.028404           x       0.030544
x       0.026357           q       0.028123           q       0.030244
                           th      0.009954           the     0.005715
                           in      0.006408           ing     0.003237
                           er      0.004755           and     0.003128
                           an      0.004352           in      0.002968
                           ou      0.003225           ed      0.002547
                           on      0.003180           to      0.002496
                           he      0.003108           of      0.002486
                           at      0.002851           en      0.001331
                           ed      0.002804           an      0.001313
                           or      0.002786           th      0.001270
                           en      0.002538           er      0.001250
                           to      0.002511           es      0.001209
                           of      0.002475           at      0.001181
                           st      0.002415           it      0.001171
                           nd      0.002297           that    0.001165
```

```
Words                                  <Nw>     quality factor
------------------------------------------------------------------
abominate                             2.0000         1.0000
achieved                              2.0000         1.0000
aemploy                               2.0000         1.0000
affrighted                            2.0000         1.0000
afternoon                             2.0000         1.0000
afterwards                            5.0000         1.0000
ahollow                               2.0000         1.0000
american                              3.0000         1.0000
anxious                               2.0000         1.0000
apartment                             2.0000         1.0000
appeared                              4.0000         1.0000
astonishment                          4.0000         1.0000
attention                             2.0000         1.0000
avenues                               2.0000         1.0000
bashful                               2.0000         1.0000
battery                               2.0000         1.0000
beefsteaks                            2.0000         1.0000
believe                               2.0000         1.0000
beloved                               2.0000         1.0000
beneath                               6.0000         1.0000
between                              12.0000         1.0000
boisterous                            3.0000         1.0000
botherwise                            2.0000         1.0000
bountiful                             2.0000         1.0000
bowsprit                              2.0000         1.0000
breakfast                             5.0000         1.0000
breeding                              2.0000         1.0000
bulkington                            3.0000         1.0000
bulwarksb                             2.0000         1.0000
bumpkin                               2.0000         1.0000
business                              6.0000         1.0000
carpenters                            2.0000         1.0000
```

summary

Enormous amount of systematic data

DNA sequence, protein sequence & structure, functional data

Various Databases (general purpose + specialized)

Analysis tools

Tasks driven by known mechanisms

Combine multiple source of data

⬇

build mechanistic models

⬇

suggest new principle