Julia Genomics Package for Bioinformatics

18.337 Final Project

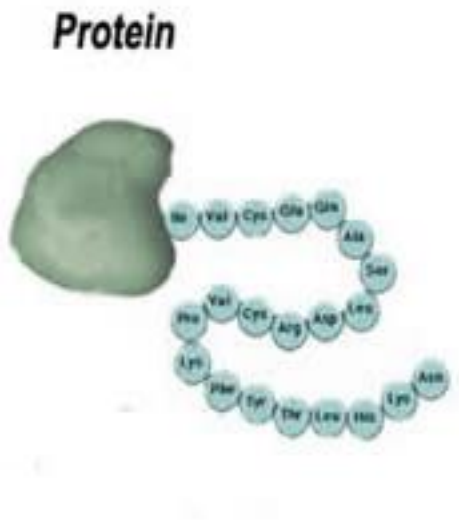
Isha Jain

December 11, 2013

Introduction | Central Dogma

Transcription

Translation



Strings in a 4-letter language:

Strings in a 4-letter language:

Strings in a 20-letter language:

Library Functions| Categories

Daily Use Toolbox

- Reverse Complement
- Translator

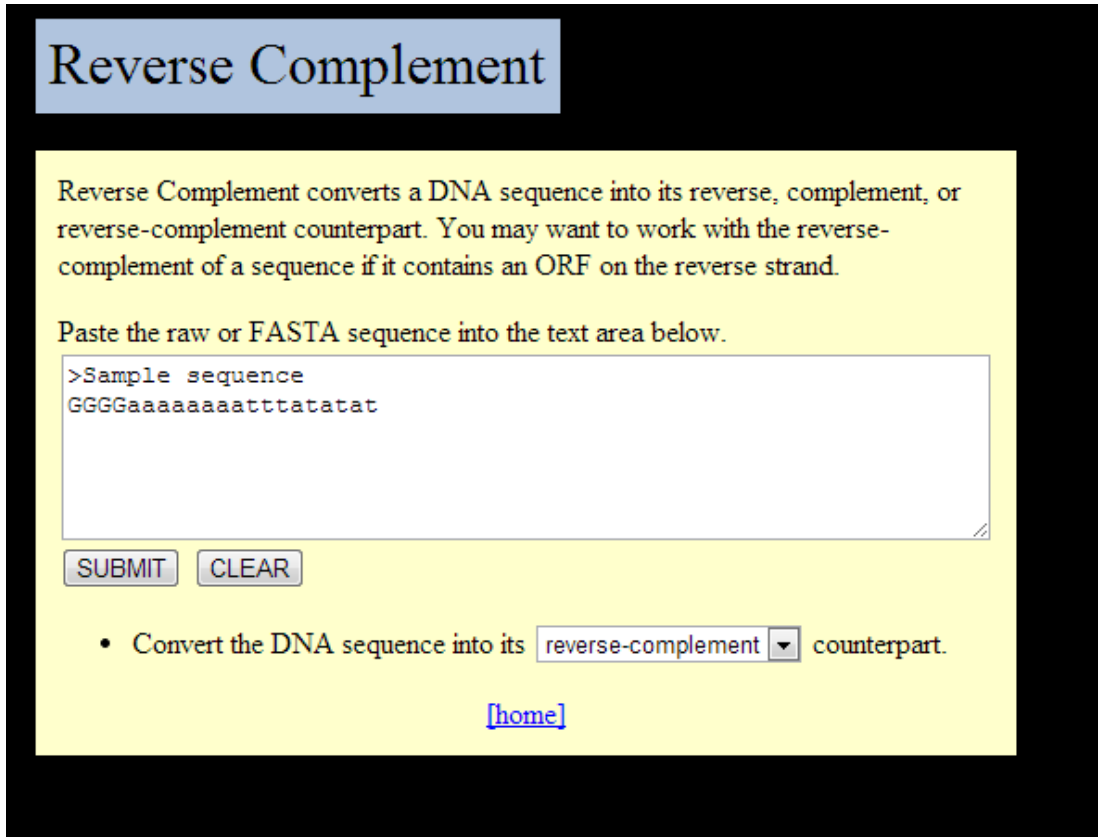
Intermediate Toolbox

- Primer Tm Calculator
- Restriction Enzyme Cutter
- Protein Atomic Composition

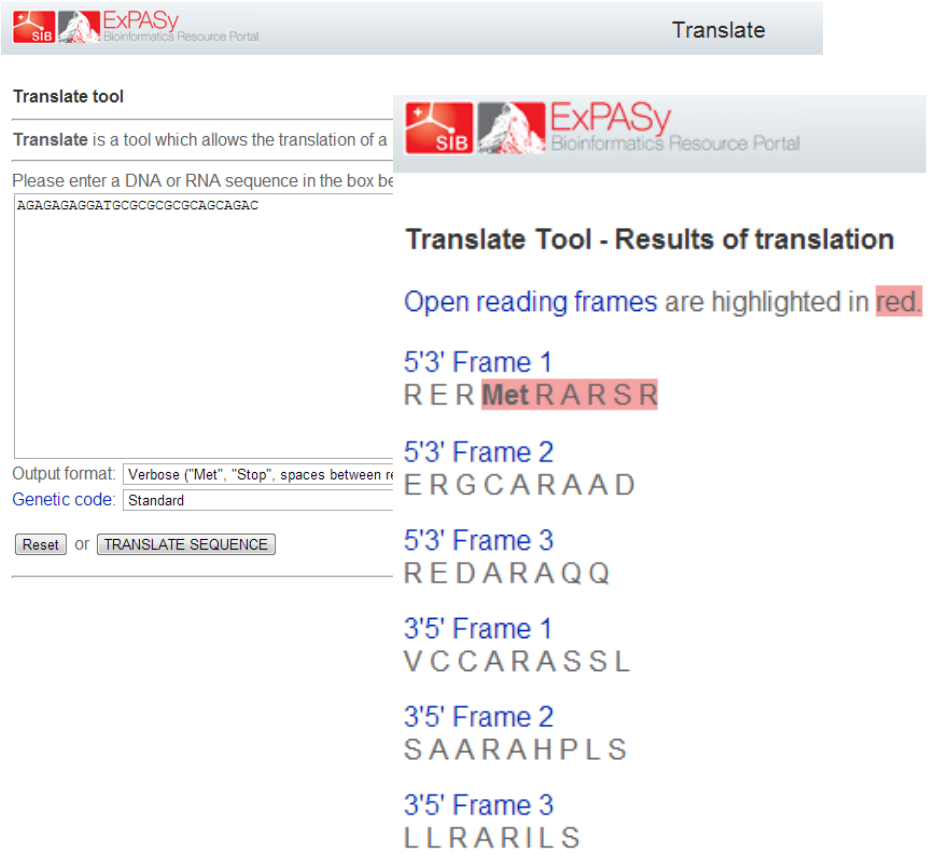
Advanced Toolbox

- Motif Finder
- Sequence Alignment
- Co-expression Analysis

Library Functions | Daily Use Functions

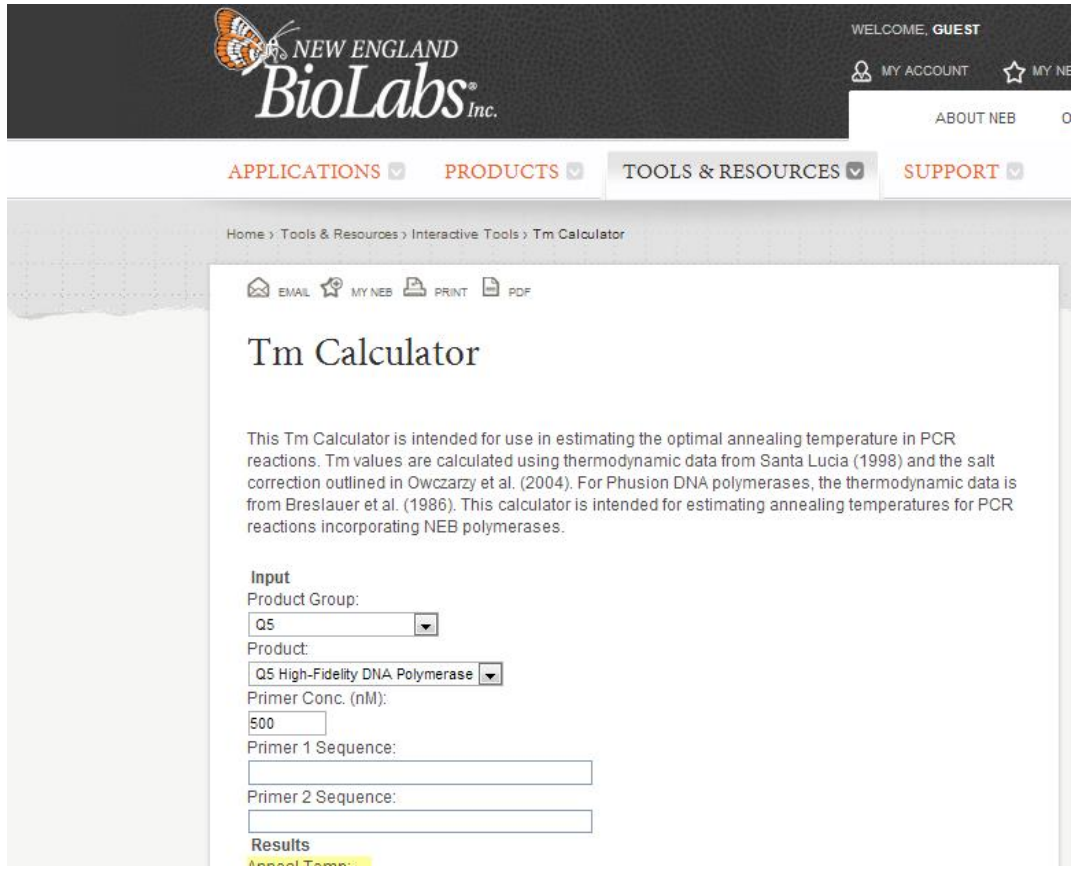


http://www.bioinformatics.org/sms/rev_comp.html



http://web.expasy.org/translate/

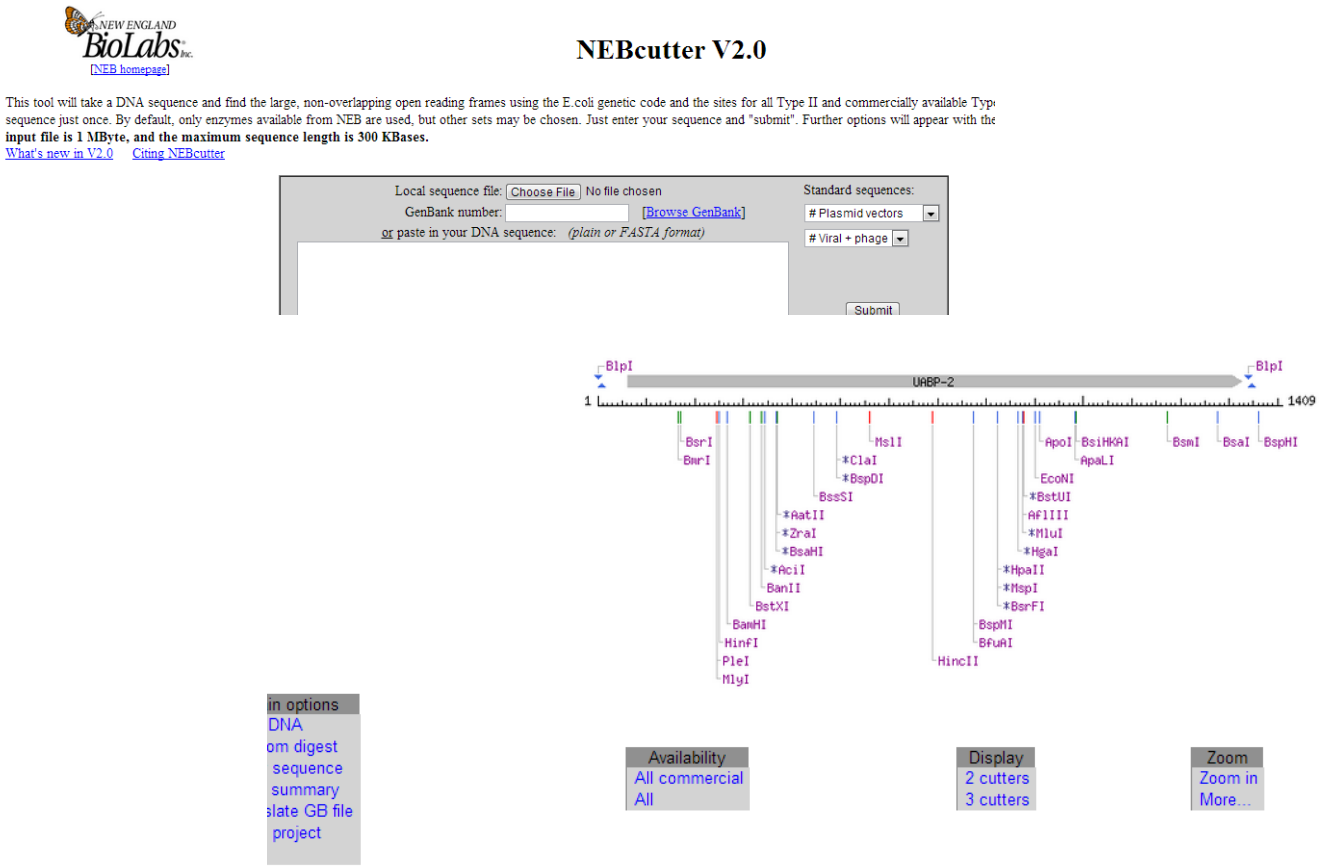
Library Functions |Intermediate Toolbox | Primer Tm



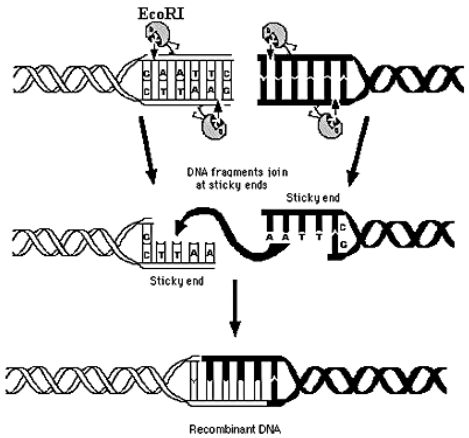
https://www.neb.com/tools-and-resources/interactive-tools/tm-calculator

- In double-stranded DNA:
  - A binds to T
  - C binds to G
- In order to amplify a region of the genome, you design short strands of DNA that bind to the sequences surrounding your piece of interest
- For amplification to occur, the short strand of DNA must have certain properties of which melting temperature is most important
- The melting temperature of a piece of DNA can be calculated based on the different types of bonds predicted

Library Functions |Intermediate Toolbox | Restriction Enzyme Cutter



http://tools.neb.com/NEBcutter2/



Restriction Enzyme Action of EcoRI

- Restriction enzymes (RE) are proteins that recognize specific sequences
- They cut the DNA with high specificity
- RE are often repurposed for molecular biology experiments

Article No. mb981947 *J. Mol. Biol.* (1998) **281**, 827–842

JMB

Extracting Regulatory Sites from the Upstream Region of Yeast Genes by Computational Analysis of Oligonucleotide Frequencies

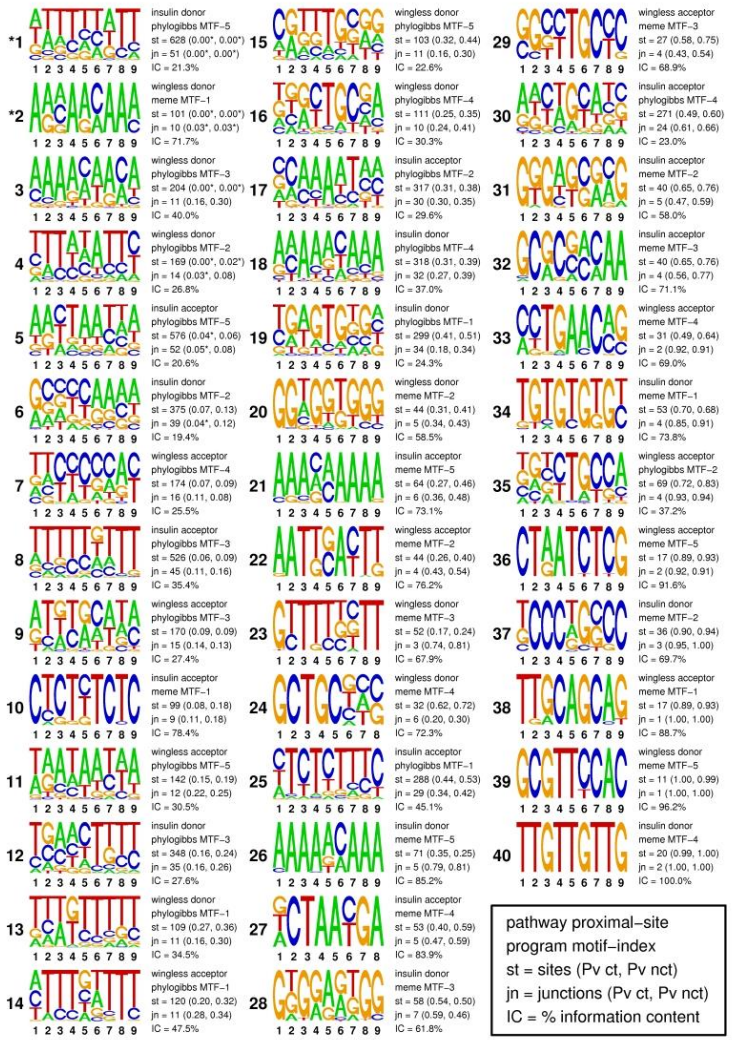
J. van Helden[1]*, B. André[2] and J. Collado-Vides[1]

[1]*Centro de Investigación sobre Fijación de Nitrógeno Universidad Nacional Autónoma de México, AP565A Cuernavaca, 62100, Morelos*

We present here a simple and fast method allowing the isolation of DNA binding sites for transcription factors from families of coregulated genes, with results illustrated in *Saccharomyces cerevisiae*. Although conceptually simple, the algorithm proved efficient for extracting, from most of the yeast regulatory families analyzed, the upstream regulatory sequences

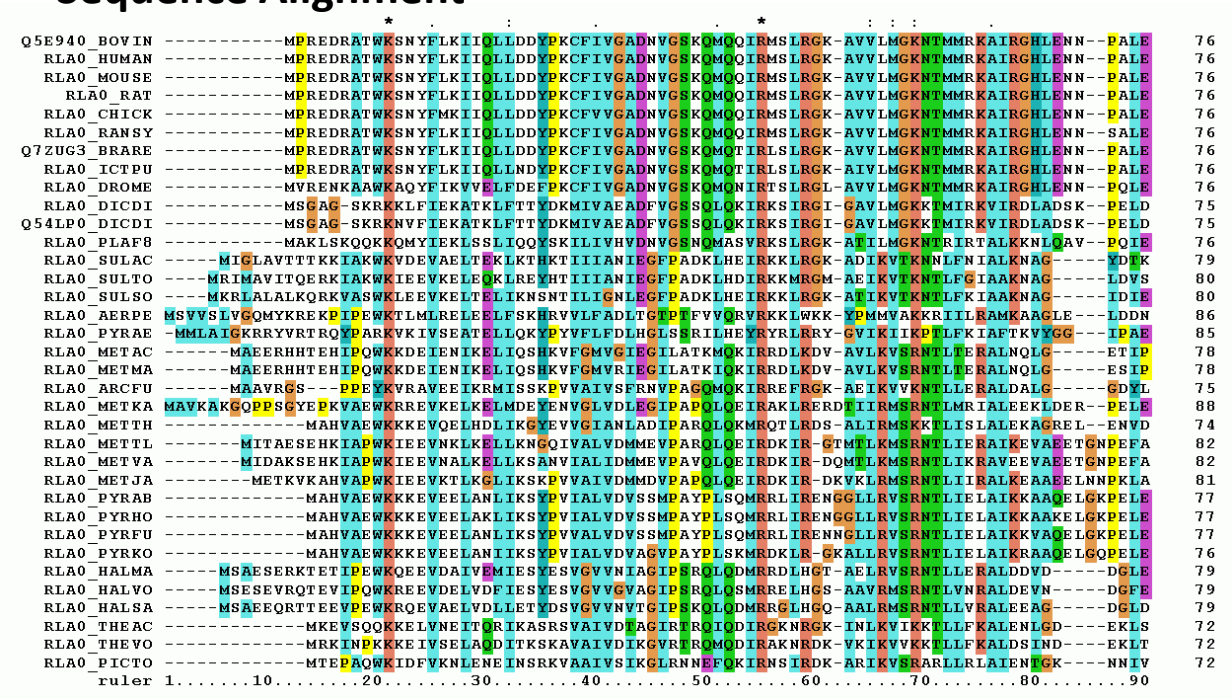
Library Functions |Advanced Toolbox | Motif Finder

1. Convert Genomic Sequence to Base 4 Representation
2. Determine distribution of all k-mers in intergenic, non-coding regions
3. Define subset of sequences that are thought to be functional
4. Determine distribution of k-mers in functional sequences
5. Use binomial to assign an enrichment score to each k-mer

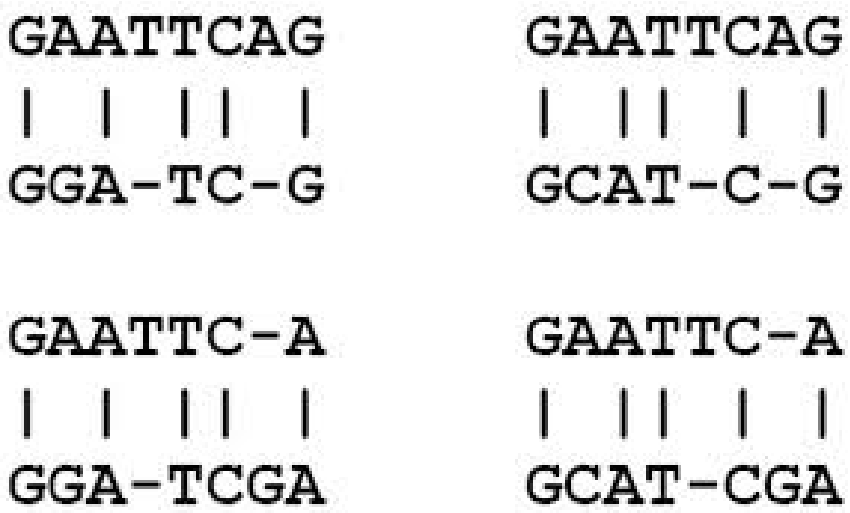


Library Functions |Advanced Toolbox | Sequence Alignment

Sequence Alignment



Simplified Representation



Needleman-Wunsch Algorithm

(A)

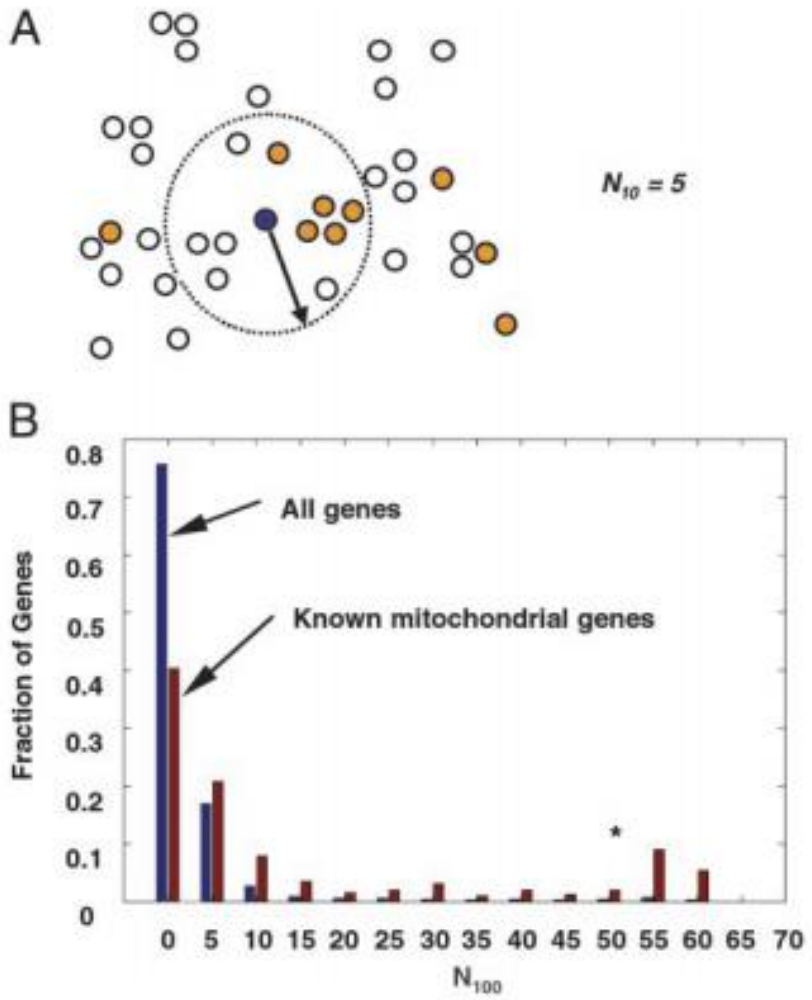
		I	S	A	L	I	G	N	E	D
	0	-4	-8	-12	-16	-20	-24	-28	-32	-36
T	-4	-1	-3	-7	-11	-15	-19	-23	-27	-31
H	-8	-5	-2	-5	-9	-13	-17	-18	-22	-26
I	-12	-4	-6	-3	-3	-5	-9	-13	-17	-21
S	-16	-8	0	-4	-5	-5	-5	-8	-12	-16
L	-20	-12	-4	-1	0	-3	-7	-8	-11	-15
I	-24	-16	-8	-5	1	4	0	-4	-8	-12
N	-28	-20	-12	-9	-3	0	4	6	2	-2
E	-32	-24	-16	-13	-7	-4	0	4	11	7

- Assign a penalty to
  - Mismatches
  - Gaps
  - Different Start Sites
- Find optimal solution for each subsequence iteratively
- After completing scoring matrix, trace back path to find the most optimal alignment
- This algorithm can be modified to have context-dependent and mismatch-specific penalties

Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics

Vamsi K. Mootha*, Pierre Lepage†, Kathleen Miller*, Jakob Bunkenborg‡, Michael Reich*, Majbrit Hjerrild‡, Terrye Delmonte*, Amelie Villeneuve†, Robert Sladek§, Fenghao Xu¶, Grant A. Mitchell‖, Charles Morin**, Matthias Mann‡, Thomas J. Hudson§, Brian Robinson¶, John D. Rioux*†‡‡, and Eric S. Lander*†‡‡§§

*Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research, Cambridge, MA 02139; †Genome Quebec Innovation Centre, McGill University, Montreal, QC, Canada H3G 1A4; ‡MDS Proteomics, 5230 Odense, Denmark; §Montreal Genome Centre, McGill University Health Centre, Montreal, QC, Canada H3G 1A4; ¶Hospital for Sick Children, Toronto, ON, Canada M5G 1X8; ‖Service de Génétique Médicale, Hôpital Sainte-Justine, Montreal, QC, Canada H3T 1C5; **Department of Pediatrics and Clinical Research Unit, Chicoutimi, QC, Canada G7H 4A3; and §§Department of Biology, Massachusetts Institute of Technology, Cambridge MA 02138



1. Input and format microarray data for 15+ experimental conditions.
2. For each gene or probeset, find the other genes/probesets that correlate well -> rank top hits.
3. Compare to gene list of interest.
4. Genes that have the largest number of neighbors that are contained within the gene list of interest are likely functionally related to the genes in the comparison list.

Ex. Application: proteome determination