

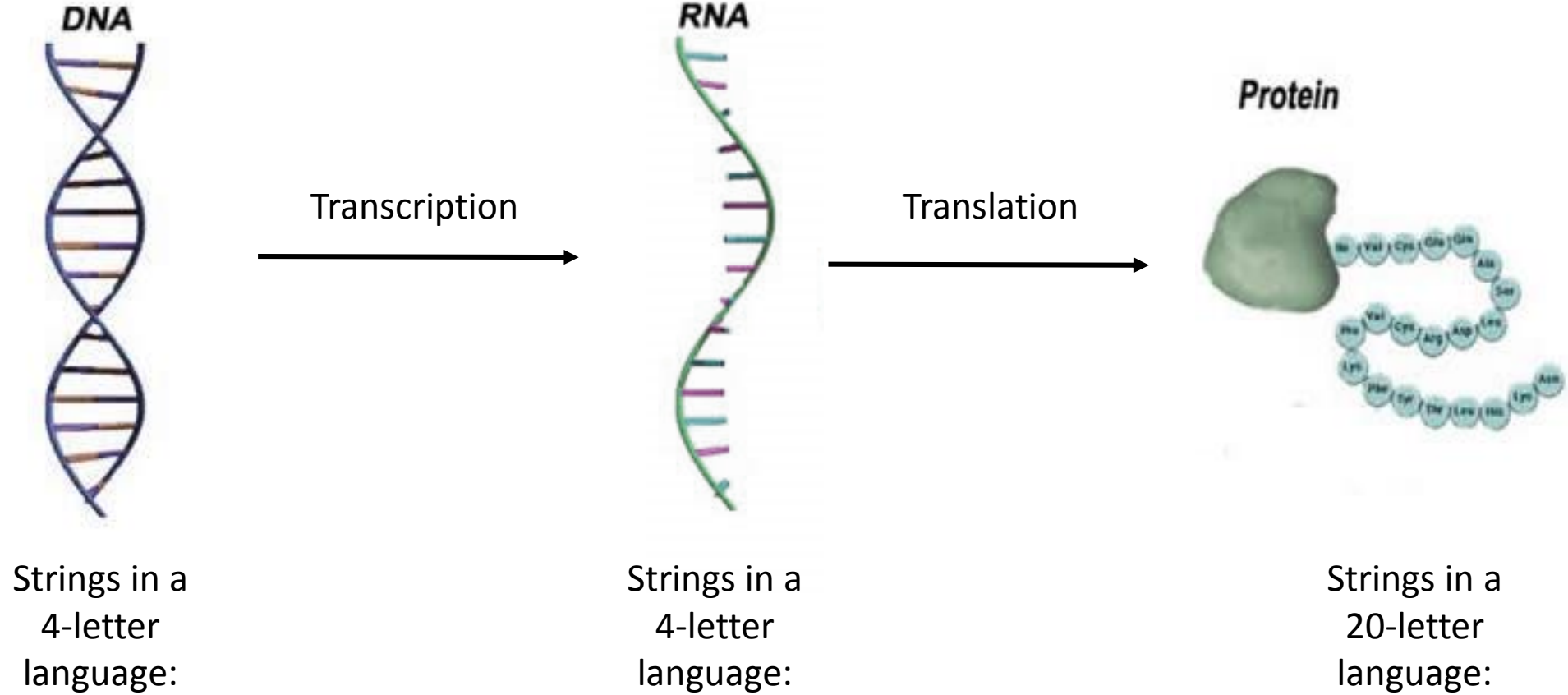


Julia Genomics Package for Bioinformatics

18.337 Final Project

Isha Jain

December 11, 2013



## 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

## Reverse Complement

Reverse Complement converts a DNA sequence into its reverse, complement, or reverse-complement counterpart. You may want to work with the reverse-complement of a sequence if it contains an ORF on the reverse strand.

Paste the raw or FASTA sequence into the text area below.

```
>Sample sequence  
GGGGaaaaaaaaatttatatat
```

SUBMIT

CLEAR

- Convert the DNA sequence into its  counterpart.

[\[home\]](#)

[http://www.bioinformatics.org/sms/rev\\_comp.html](http://www.bioinformatics.org/sms/rev_comp.html)

Translate

Translate tool

Translate is a tool which allows the translation of a DNA or RNA sequence

Please enter a DNA or RNA sequence in the box below

```
AGAGAGAGGATGCGCGCGCAGCAGAC
```

**Translate Tool - Results of translation**

Open reading frames are highlighted in red.

5'3' Frame 1  
RER **Met** RARSR

5'3' Frame 2  
ERGCARAAD

5'3' Frame 3  
REDARAQQ

3'5' Frame 1  
VCCARASSL

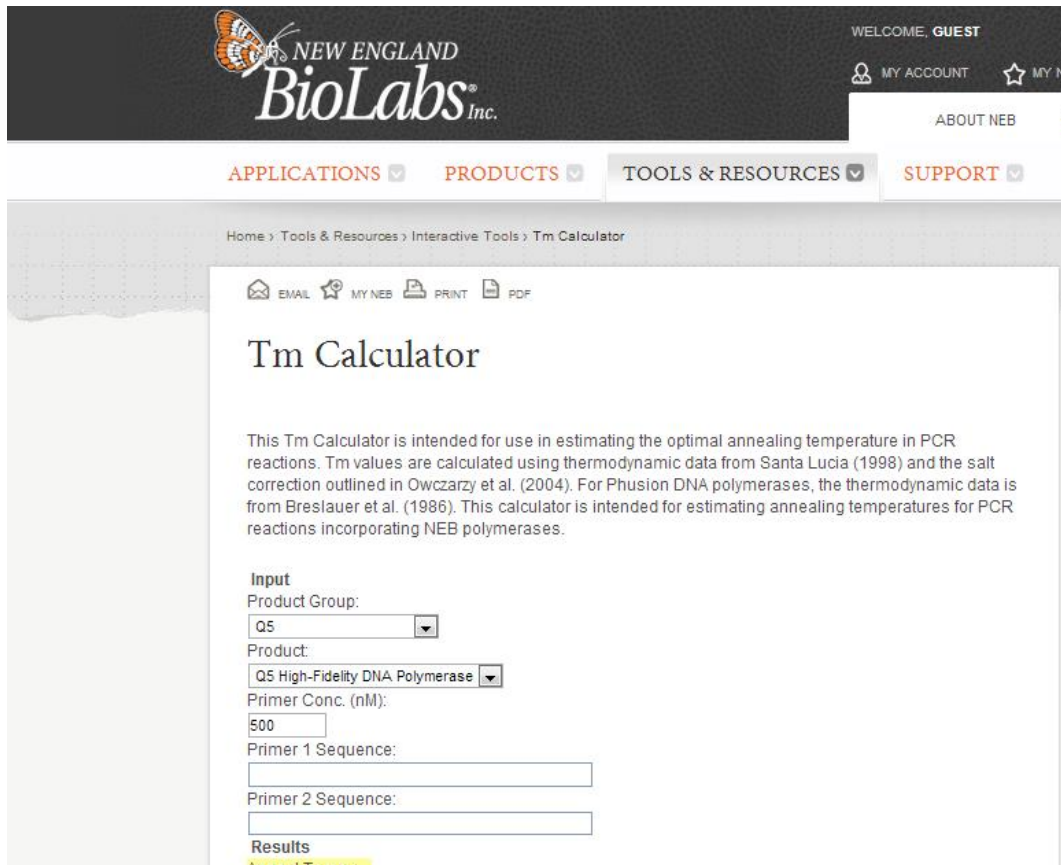
3'5' Frame 2  
SAARAHPLS

3'5' Frame 3  
LLRARILS

Output format:  Genetic code:

or

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



The screenshot shows the NEB website's Tm Calculator page. The header includes the NEB logo and navigation links like 'APPLICATIONS', 'PRODUCTS', 'TOOLS & RESOURCES', and 'SUPPORT'. The main content area is titled 'Tm Calculator' and contains a descriptive paragraph about the tool's purpose and the scientific data it uses. Below the text is an 'Input' section with several fields: 'Product Group' (dropdown menu), 'Product' (dropdown menu), 'Primer Conc. (nM)' (text input), 'Primer 1 Sequence' (text input), and 'Primer 2 Sequence' (text input). A 'Results' section is partially visible at the bottom.

<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

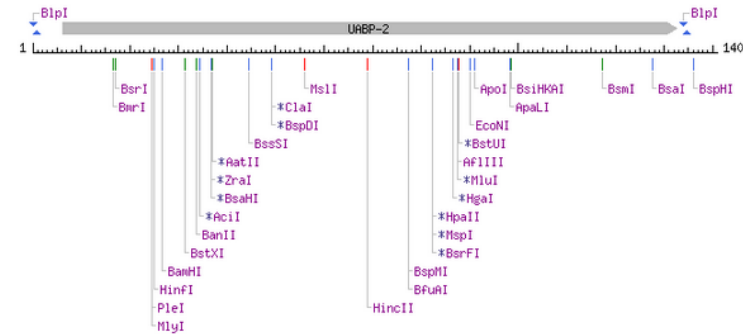


## NEBcutter V2.0

This tool will take a DNA sequence and find the large, non-overlapping open reading frames using the E. coli genetic code and the sites for all Type II and commercially available Type III restriction enzymes just once. By default, only enzymes available from NEB are used, but other sets may be chosen. Just enter your sequence and "submit". Further options will appear with the input file is 1 MByte, and the maximum sequence length is 300 KBases.

[What's new in V2.0](#) [Using NEBcutter](#)

Local sequence file:  No file chosen  
GenBank number:    
or paste in your DNA sequence: (plain or FASTA format)  
  
Standard sequences:  
# Plasmid vectors   
# Viral + phage

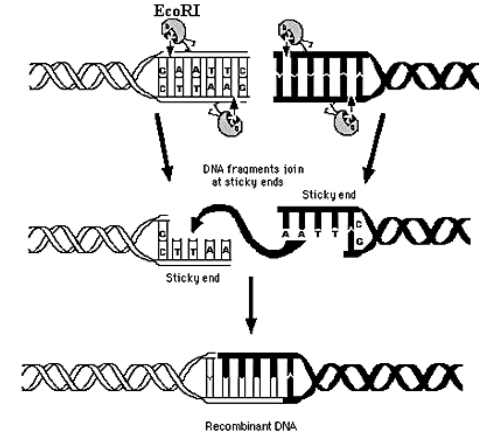


- in options
- DNA
- om digest
- sequence
- summary
- slate GB file
- project

Availability  
All commercial  
All

Display  
2 cutters  
3 cutters

Zoom  
Zoom in  
More...



### 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

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

---

**JMIB**



---

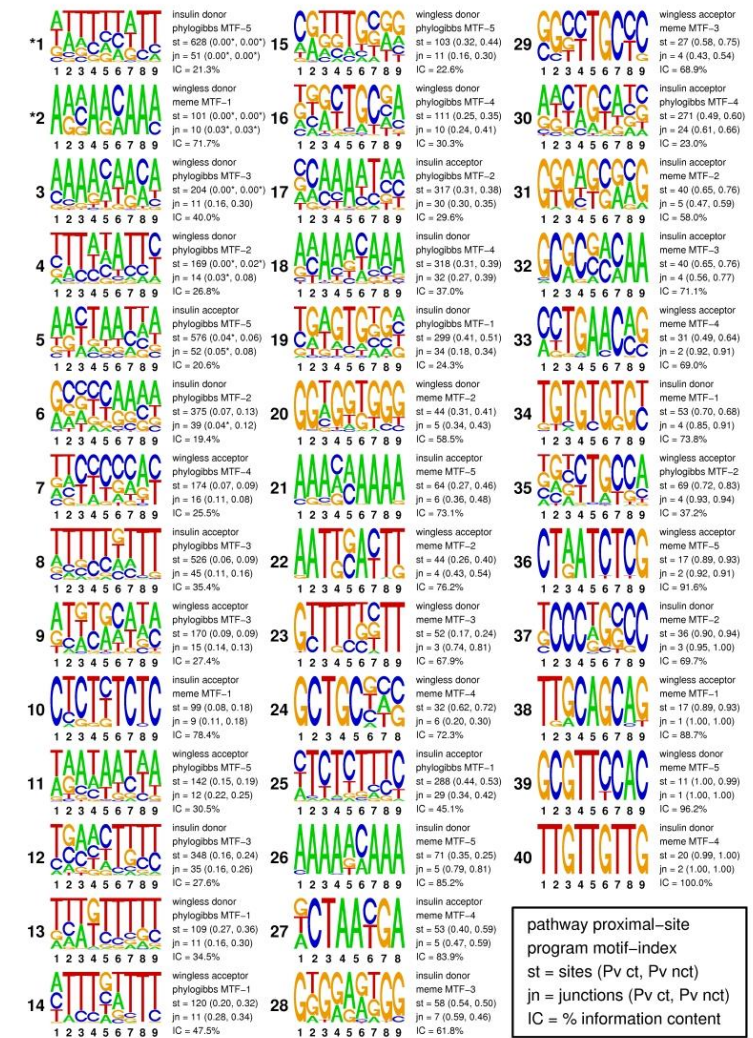
## **Extracting Regulatory Sites from the Upstream Region of Yeast Genes by Computational Analysis of Oligonucleotide Frequencies**

**J. van Helden<sup>1\*</sup>, B. André<sup>2</sup> and J. Collado-Vides<sup>1</sup>**

<sup>1</sup>*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

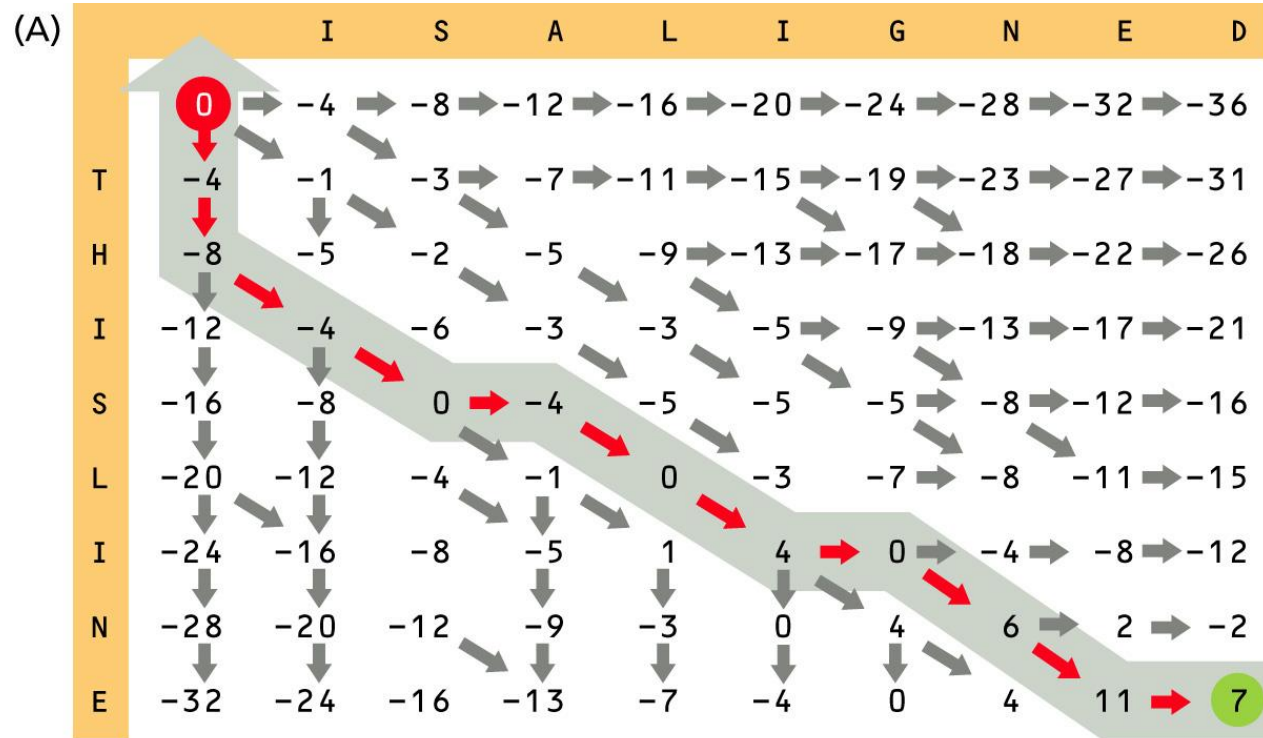
1. Convert Genomic Sequence to Base 4 Representation
2. Determine distribution of all k-mers in intergenic, non-coding regions
3. Define subset of sequences that are thought to be functional
4. Determine distribution of k-mers in functional sequences
5. Use binomial to assign an enrichment score to each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







## Needleman-Wunsch Algorithm



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

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

