Challenging Problems in Bioinformatics and Computational Biology

Yosi Shibberu
Jimma University

October 20, 2014

Applications

- disease diagnosis

Applications

- disease diagnosis
- vaccines and drug development

Applications

- disease diagnosis
- vaccines and drug development
- cataloging bio-diversity

Applications

- disease diagnosis
- vaccines and drug development
- cataloging bio-diversity
- seed development and certification

Agriculture

- Population growth, climate change and water shortages threaten world food security.

Agriculture

- Population growth, climate change and water shortages threaten world food security.
- As the climate changes, plants are subject to new pests and plant diseases and don’t have time to evolve new defenses.

Agriculture[1]

- Bioinformatics can be used to:

[1] http://www.csa.com/discoveryguides/gmfood/overview.php

Agriculture[1]

- Bioinformatics can be used to:
  - identify mechanisms of drought and pest resistance.

[1] http://www.csa.com/discoveryguides/gmfood/overview.php

Agriculture[1]

- Bioinformatics can be used to:
  - identify mechanisms of drought and pest resistance.
  - develop GM (genetically-modified) crops.

[1] http://www.csa.com/discoveryguides/gmfood/overview.php

Agriculture[1]

- Bioinformatics can be used to:
  - identify mechanisms of drought and pest resistance.
  - develop GM (genetically-modified) crops.
- Benefits:

[1] http://www.csa.com/discoveryguides/gmfood/overview.php

Agriculture[1]

- Bioinformatics can be used to:
    - identify mechanisms of drought and pest resistance.
    - develop GM (genetically-modified) crops.
- Benefits:
    - drought/salinity tolerance

[1] http://www.csa.com/discoveryguides/gmfood/overview.php

Agriculture[1]

- Bioinformatics can be used to:
  - identify mechanisms of drought and pest resistance.
  - develop GM (genetically-modified) crops.
- Benefits:
  - drought/salinity tolerance
  - heat/cold tolerence

[1] http://www.csa.com/discoveryguides/gmfood/overview.php

Agriculture[1]

- Bioinformatics can be used to:
  - identify mechanisms of drought and pest resistance.
  - develop GM (genetically-modified) crops.
- Benefits:
  - drought/salinity tolerance
  - heat/cold tolerence
  - pest/disease resistance

[1] http://www.csa.com/discoveryguides/gmfood/overview.php

Agriculture[1]

- Bioinformatics can be used to:
  - identify mechanisms of drought and pest resistance.
  - develop GM (genetically-modified) crops.
- Benefits:
  - drought/salinity tolerance
  - heat/cold tolerence
  - pest/disease resistance
  - herbicide tolerance

[1] http://www.csa.com/discoveryguides/gmfood/overview.php

Agriculture[1]

- Bioinformatics can be used to:
  - identify mechanisms of drought and pest resistance.
  - develop GM (genetically-modified) crops.
- Benefits:
  - drought/salinity tolerance
  - heat/cold tolerence
  - pest/disease resistance
  - herbicide tolerance
  - enhanced nutrition

[1] http://www.csa.com/discoveryguides/gmfood/overview.php

Agriculture[1]

- Bioinformatics can be used to:
  - identify mechanisms of drought and pest resistance.
  - develop GM (genetically-modified) crops.
- Benefits:
  - drought/salinity tolerance
  - heat/cold tolerence
  - pest/disease resistance
  - herbicide tolerance
  - enhanced nutrition
- Dangers:

[1] http://www.csa.com/discoveryguides/gmfood/overview.php

Agriculture[1]

- Bioinformatics can be used to:
  - identify mechanisms of drought and pest resistance.
  - develop GM (genetically-modified) crops.
- Benefits:
  - drought/salinity tolerance
  - heat/cold tolerence
  - pest/disease resistance
  - herbicide tolerance
  - enhanced nutrition
- Dangers:
  - Unknown/unintended impact on the natural environment.

[1] http://www.csa.com/discoveryguides/gmfood/overview.php

Medicine

- Bioinformatics is used to:

Medicine

- Bioinformatics is used to:
  - guide the treatment of cancers.

Medicine

- Bioinformatics is used to:
  - guide the treatment of cancers.
  - identify drug targets and lead compounds for drug development.

Medicine

- Bioinformatics is used to:
  - guide the treatment of cancers.
  - identify drug targets and lead compounds for drug development.
  - warn parents of potential genetic defects.

Medicine

- Bioinformatics is used to:
  - guide the treatment of cancers.
  - identify drug targets and lead compounds for drug development.
  - warn parents of potential genetic defects.
  - diagnose certain diseases.

Central Dogma of Biology[2]

[2] Francis Crick, 1958.

Central Dogma of Biology[2]

DNA

[2] Francis Crick, 1958.

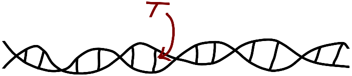
Central Dogma of Biology[2]

DNA

[2] Francis Crick, 1958.

Central Dogma of Biology[2]

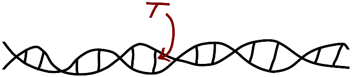
DNA



[2] Francis Crick, 1958.

Central Dogma of Biology[2]

DNA

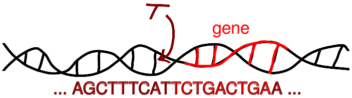


... AGCTTTCATTCTGACTGAA ...

[2] Francis Crick, 1958.

Central Dogma of Biology[2]

DNA

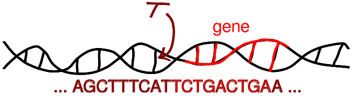


[2] Francis Crick, 1958.

Central Dogma of Biology[2]

DNA

transcription



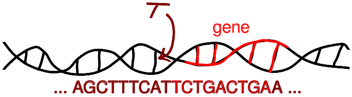
[2] Francis Crick, 1958.

Central Dogma of Biology[2]

DNA

transcription

RNA



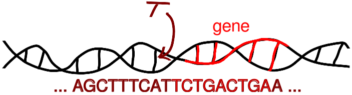
[2] Francis Crick, 1958.

Central Dogma of Biology[2]

DNA

transcription

RNA



[2] Francis Crick, 1958.

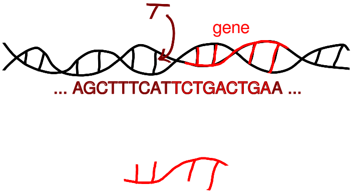
Central Dogma of Biology[2]

DNA

transcription

RNA

translation



[2] Francis Crick, 1958.

Central Dogma of Biology[2]

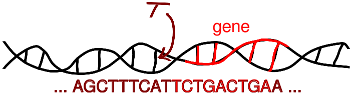
DNA

transcription

RNA

translation

Protein



[2] Francis Crick, 1958.

Central Dogma of Biology[2]

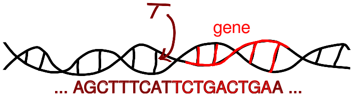
DNA

transcription

RNA

translation

Protein



[2] Francis Crick, 1958.

Central Dogma of Biology[2]

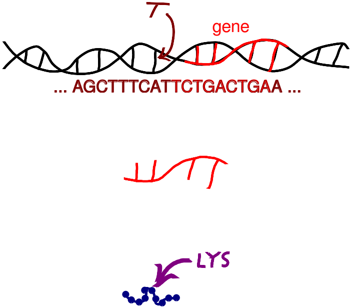
DNA

transcription

RNA

translation

Protein



[2] Francis Crick, 1958.

Central Dogma of Biology[2]

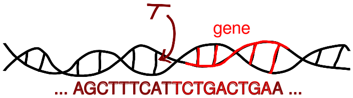
DNA

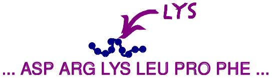
transcription

RNA

translation

Protein





[2] Francis Crick, 1958.

Central Dogma of Biology[2]

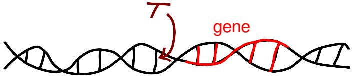
DNA

transcription

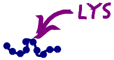
RNA

translation

Protein



... AGCTTTCATTCTGACTGAA ...



... ASP ARG LYS LEU PRO PHE ...

[2] Francis Crick, 1958.

Central Dogma of Biology[2]

DNA

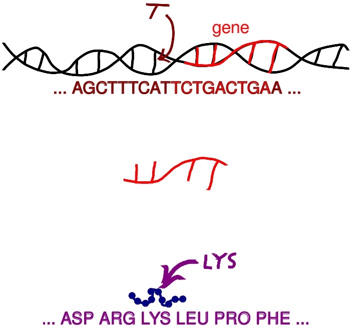
transcription

RNA

translation

Protein

folding



[2] Francis Crick, 1958.

Central Dogma of Biology[2]

DNA

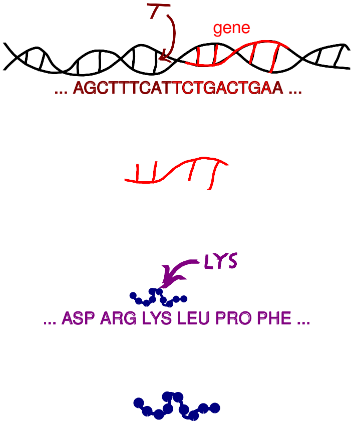
transcription

RNA

translation

Protein

folding



[2] Francis Crick, 1958.

Central Dogma of Biology[2]

DNA

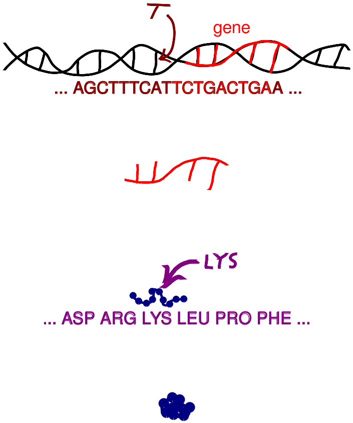
transcription

RNA

translation

Protein

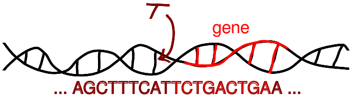
folding



[2] Francis Crick, 1958.

Central Dogma of Biology[2]

DNA

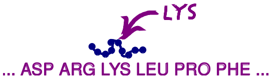


transcription

RNA

translation

Protein

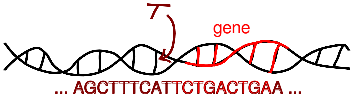


folding dynamics

[2] Francis Crick, 1958.

Central Dogma of Biology[2]

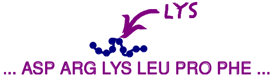
DNA



transcription

RNA

translation



Protein

---

folding
dynamics

[2] Francis Crick, 1958.

Central Dogma of Biology[2]

DNA

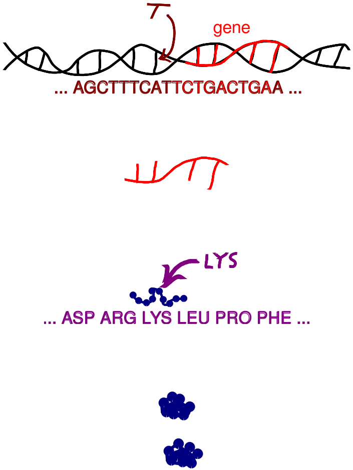
transcription

RNA

translation

Protein

folding
dynamics



[2] Francis Crick, 1958.

Central Dogma of Biology[2]

DNA

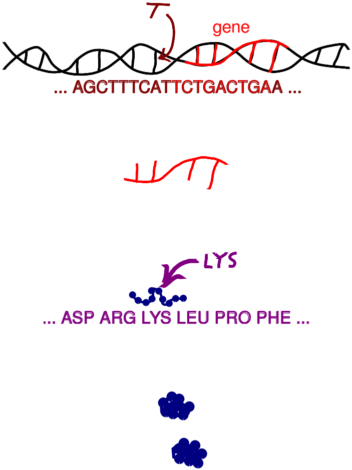
transcription

RNA

translation

Protein

folding
dynamics



[2] Francis Crick, 1958.

Central Dogma of Biology[2]

DNA

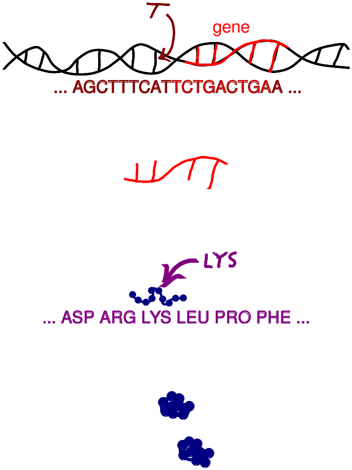
transcription

RNA

translation

Protein

folding
dynamics



[2] Francis Crick, 1958.

Central Dogma of Biology[2]

DNA

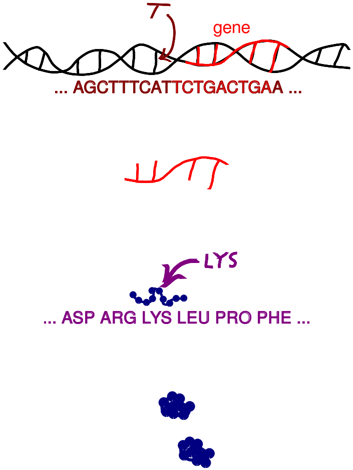
transcription

RNA

translation

Protein

folding
dynamics
function



[2] Francis Crick, 1958.

Central Dogma of Biology[2]

DNA

transcription

RNA

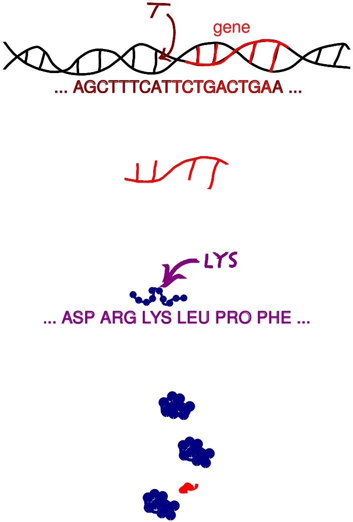
translation

Protein

folding

dynamics

function



gene

... AGCTTTCATTCTGACTGAA ...

... ASP ARG LYS LEU PRO PHE ...

[2] Francis Crick, 1958.

Central Dogma of Biology[2]

DNA

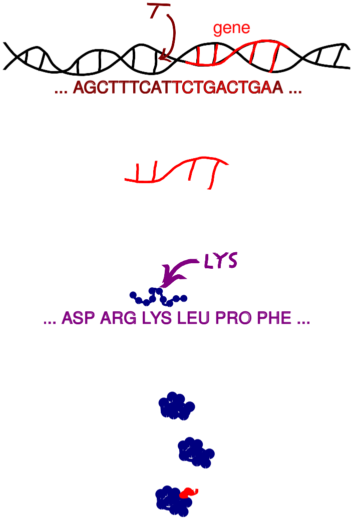
transcription

RNA

translation

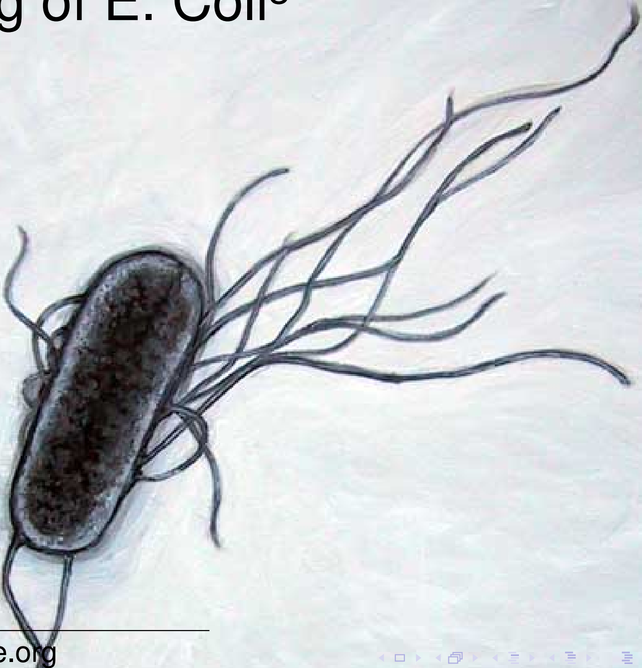
Protein

folding
dynamics
function



[2] Francis Crick, 1958.

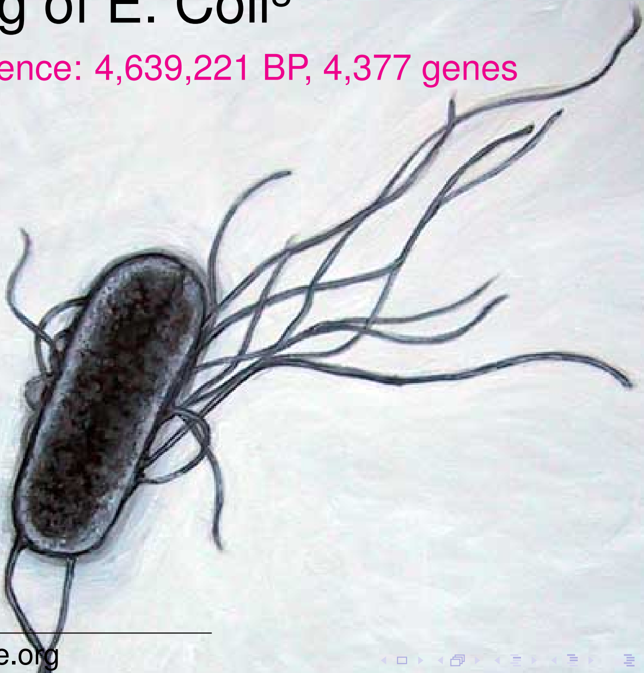
Painting of E. Coli[3]



[3] shardcore.org

Painting of E. Coli[3]

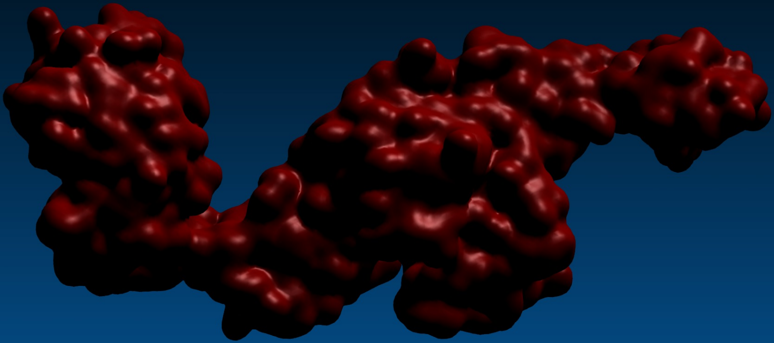
DNA Sequence: 4,639,221 BP, 4,377 genes

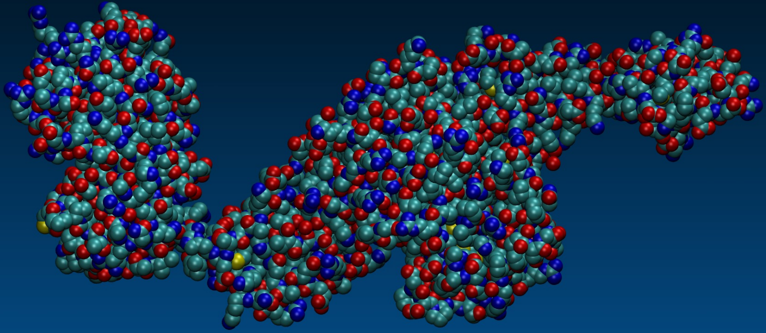


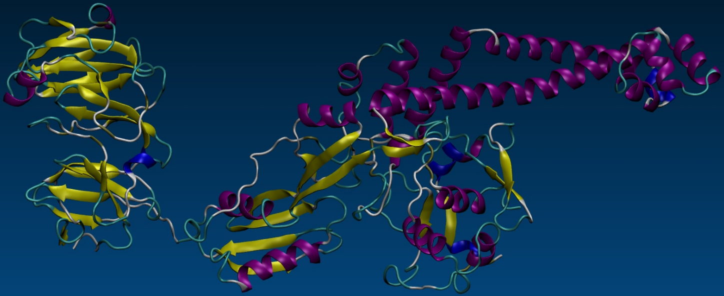
[3] shardcore.org

E. Coli parC Gene

- 716 residues
- 5,367 atoms

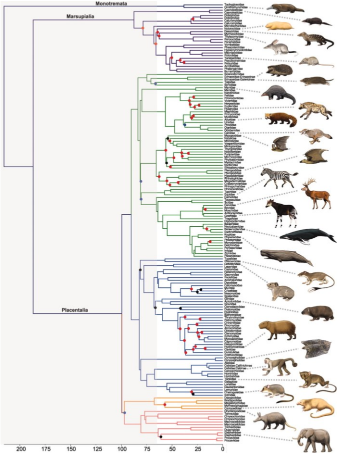






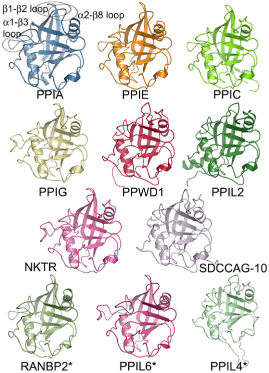
Evolution

Biology



Mammal tree using 26 genes.[4]

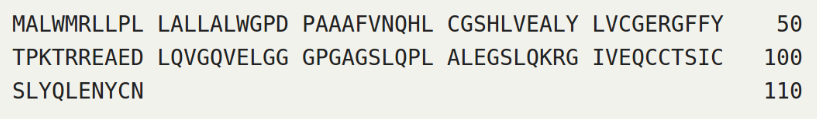
Biochemistry



Human cyclophilins protein f

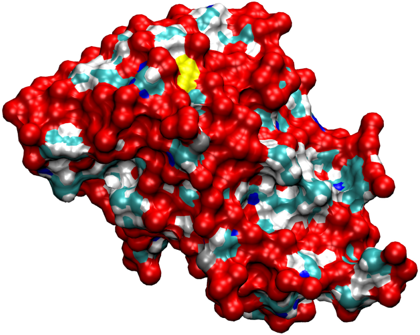
[4] Meredith, R. W. et al. (2011). Impacts of the Cretaceious terrestrial revolution and KPg extinction on mammal diversification. Science 334:521-524.

Insulin Protein Sequence



```
MALWMRLLPL LALLALWGPD PAAAFVNQHL CGSHLVEALY LVCGERGFFY    50
TPKTRREAED LQVGQVELGG GPGAGSLQPL ALEGSLQKRG IVEQCCTSIC   100
SLYQLENYCN                                               110
```

Insulin Protein Structure

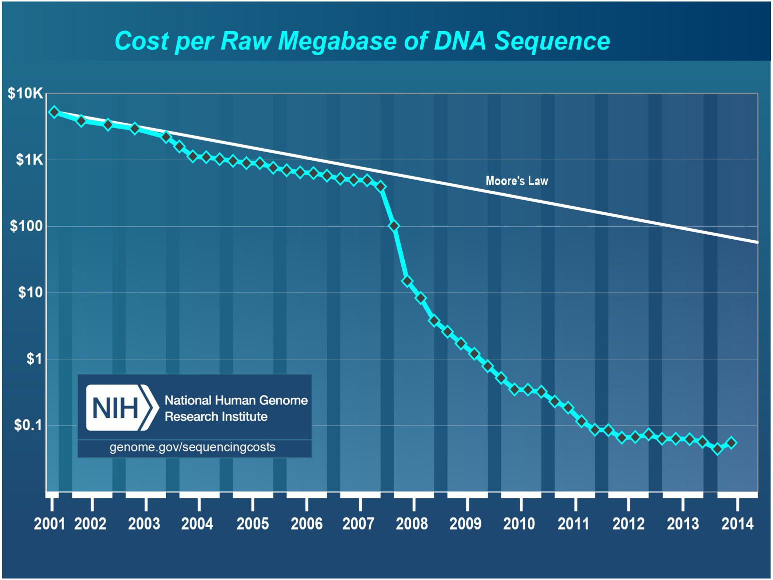


Insulin Sequence for 11 Species

```
       dog P01321 INS_CANFA 1   MALWMRLLPLLALLALWAPAPTRAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREVED   60
   hamster P01313 INS_CRILO 1   MTLWMRLLPLLTLLVLWEPNPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKSRRGVED   60
       cat P06306 INS_FELCA 1   MAPWTRLLPLLALLSLWIPAPTRAFVNQHLCGSHLVEALYLVCGERGFFYTPKARREAED   60
   gorilla Q6YK33 INS_GORGO 1   MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED   60
     human P01308 INS_HUMAN 1   MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED   60
    monkey P30406 INS_MACFA 1   MALWMRLLPLLALLALWGPDPAPAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED   60
chimpanzee P30410 INS_PANTR 1   MALWMRLLPLLVLLALWGPDPASAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED   60
 orangutan Q8HXV2 INS_PONPY 1   MALWMRLLPLLALLALWGPDPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED   60
       rat Q62587 INS_PSAOB 1   MALWMRLLPLLAFLILWEPSPAHAFVNQHLCGSHLVEALYLVCGERGFFYTPKFRRGVDD   60
    rabbit P01311 INS_RABIT 1   MASLAALLPLLALLVLCRLDPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKSRREVEE   60
  squirrel Q91XI3 INS_SPETR 1   MALWTRLLPLLALLALLGPDPAQAFVNQHLCGSHLVEALYLVCGERGFFYTPKSRREVEE   60
                                *:    *****.:* *    *: ****************************** **  .::

       dog P01321 INS_CANFA 61  LQVRDVELAGAPGEGGLQPLALEGALQKRGIVEQCCTSICSLYQLENYCN  110
   hamster P01313 INS_CRILO 61  PQVAQLELGGGPGADDLQTLALEVAQQKRGIVDQCCTSICSLYQLENYCN  110
       cat P06306 INS_FELCA 61  LQGKDAELGEAPGAGGLQPSALEAPLQKRGIVEQCCASVCSLYQLEHYCN  110
   gorilla Q6YK33 INS_GORGO 61  LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  110
     human P01308 INS_HUMAN 61  LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  110
    monkey P30406 INS_MACFA 61  PQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  110
chimpanzee P30410 INS_PANTR 61  LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  110
 orangutan Q8HXV2 INS_PONPY 61  LQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN  110
       rat Q62587 INS_PSAOB 61  PQMPQLELGGSPGAGDLRALALEVARQKRGIVEQCCTGICSLYQLENYCN  110
    rabbit P01311 INS_RABIT 61  LQVGQAELGGGPGAGGLQPSALELALQKRGIVEQCCTSICSLYQLENYCN  110
  squirrel Q91XI3 INS_SPETR 61  QQGGQVELGGGPGAGLPQPLALEMALQKRGIVEQCCTSICSLYQLENYCN  110
                                 *  : **. .**    :  ***   ******:***:.:*******.***
```

DNA Sequencing Costs[6]



[6] National Human Genome Research Institute

Illumina Sequencers

- known for speed and accuracy.

Illumina Sequencers

- known for speed and accuracy.
- reduced cost to sequence human genomes from 3 billon to 1 thousand US.

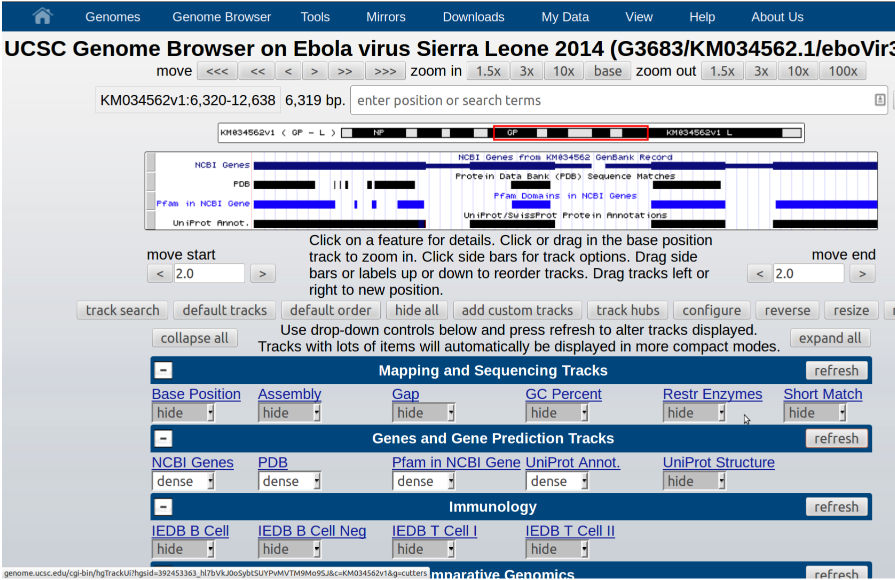
Illumina Sequencers

- known for speed and accuracy.
- reduced cost to sequence human genomes from 3 billon to 1 thousand US.
- has currently sequenced 90% of all known DNA sequences.

Illumina Sequencers

- known for speed and accuracy.
- reduced cost to sequence human genomes from 3 billon to 1 thousand US.
- has currently sequenced 90% of all known DNA sequences.
- top-of-the-line machine costs approx. 1 million US.

Ebola Web Browser



UCSC Genome Browser on Ebola virus Sierra Leone 2014 (G3683/KM034562.1/eboVir3

KM034562v1:6,320-12,638 6,319 bp. enter position or search terms

KM034562v1 (GP - L)
NP
GP
KM034562v1 L
NCBI Genes
PDB
Pfam in NCBI Gene
UniProt Annot.
NCBI Genes from KM034562 GenBank Record
Protein Data Bank (PDB) Sequence Matches
Pfam Domains in NCBI Genes
UniProt/SwissProt Protein Annotations

move start < 2.0 >

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

move end < 2.0 >

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

Mapping and Sequencing Tracks refresh

Base Position	Assembly	Gap	GC Percent	Restr Enzymes	Short Match
hide	hide	hide	hide	hide	hide

Genes and Gene Prediction Tracks refresh

NCBI Genes	PDB	Pfam in NCBI Gene	UniProt Annot.	UniProt Structure
dense	dense	dense	dense	hide

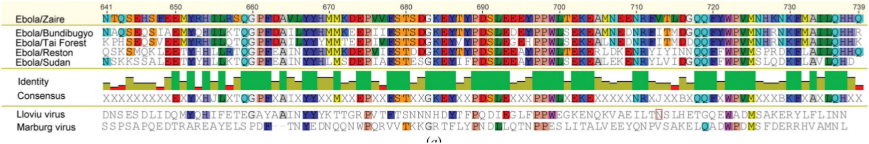
Immunology refresh

IEDB B Cell	IEDB B Cell Neg	IEDB T Cell I	IEDB T Cell II
hide	hide	hide	hide

genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=392453363_hl7bVkJ0oSybtSUYPvMVTM9Mo9SJ&c=KM034562v1&g=cutters

mparative Genomics refresh

An Ebola Protein Sequence[7]


641
650
660
670
680
690
700
710
720
730
739
Ebola/Zaire NTQSEHSFEEMYRHILRSQGPFDAVLYYHMMKDEPVVFSTSDGKEYTYPDSLEEEYPPWLTEKEAMNEENRFVTLDGQQFYWPVMNHKNKFMAILQHHQ
Ebola/Bundibugyo NAQSEQSIAEMYQHILKTQGPFDAILYYHMMKEEPIIFSTSDGKEYTYPDSLEDEYPPWLSEKEAMNEDNRFITMDGQQFYWPVMNHRNKFMAILQHHR
Ebola/Tai Forest KPHSEQSVEEMYRHILQTQGPFDAILYYYMMTEEPIVFSTSDGKEYVYPDSLEGEHPPWLSEKEALNEDNRFITMDDQQFYWPVMNHRNKFMAILQHHK
Ebola/Reston QSKSMQKLEETYHHLLRTQGPFEAINYYHMMKDEPVIFSTDDGKEYTYPDSLEEAYPPWLTEKERLDKENRYIYINNQQFFWPVMSPRDKFLAILQHHQ
Ebola/Sudan NSKKSSALEETYYHLLKTQGPFEAINYYHLMSDEPIAFSTESGKEYIFPDSLEEAYPPWLSEKEALEKENRYLVIDGQQFFWPVMSLQDKFLAVLQHD
Identity
Consensus XXXXXXXXXEXYXHJLXTQGPFXAIXYYXXMXXEPXXFSTXXGKEYXXPDSLEXXXPPWLXEKEXXXXXNRXJXXBXQQFXWPVMXXXBKFXAXLQHXX
Lloviu virus DNSESDLIDQMYQHIFETEGAYAAINYYYKTTGRPVTFTSNNNHDYTFPQDIEGLFPPWEGKENQKVAEILTNSLHETGQEWADMSAKERYLFLINN
Marburg virus SSPSAPQEDTRAREAYELSPDF--TNYEDNQQNWPQRVVTKKGRTFLYPNDLLQTNPPESLITALVEEYQNPVSAKELQADWPDMSFDERRHVAMNL

(a)

[7] Dziubanska, et al, 2014.

An Ebola Protein Structure[8]

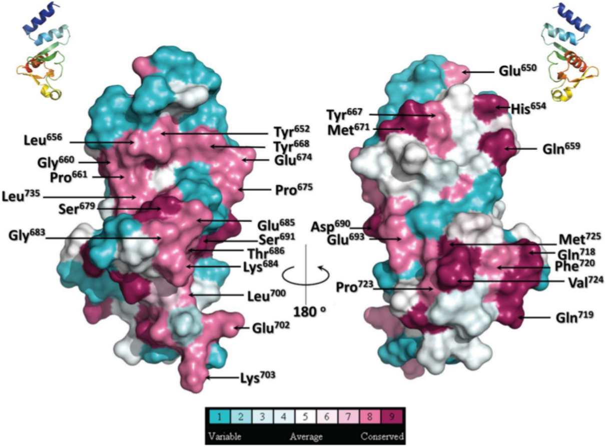
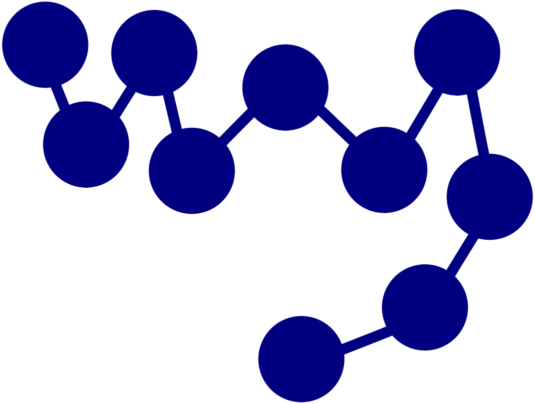


Figure 9
Graphical representation of the surface amino-acid conservation using the crystal structure of the Zaire EBOV NP^{Ct}. The color scale is based upon the level of conservation as determined by the *ConSurf* server. Categories 8 and 9 correspond to fully conserved residues. Small ribbon diagrams are shown at the top for the viewer's convenience.

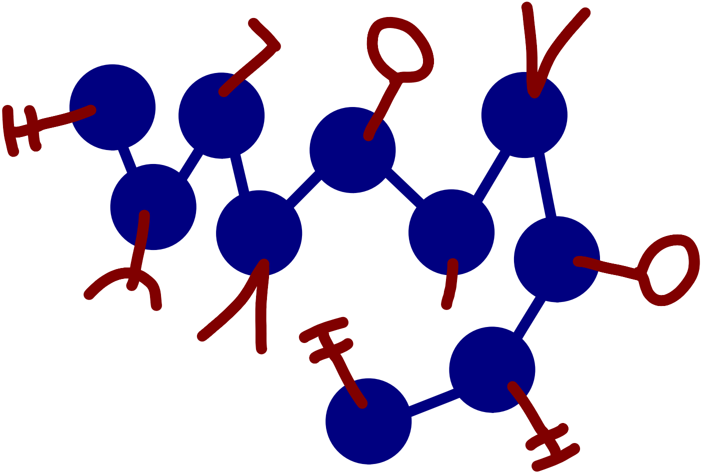
[8] Dziubanska, et al, 2014.

Protein Folding

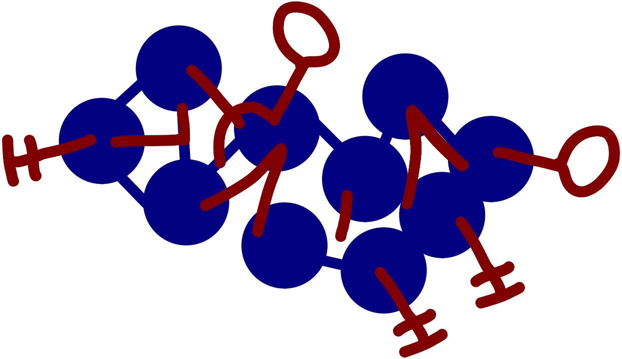
Protein Folding

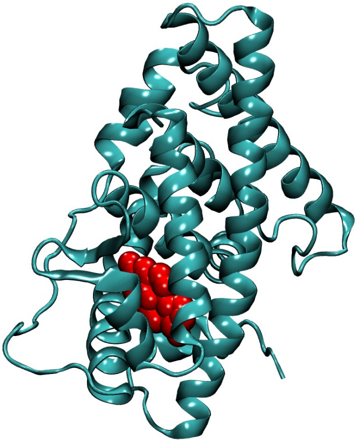


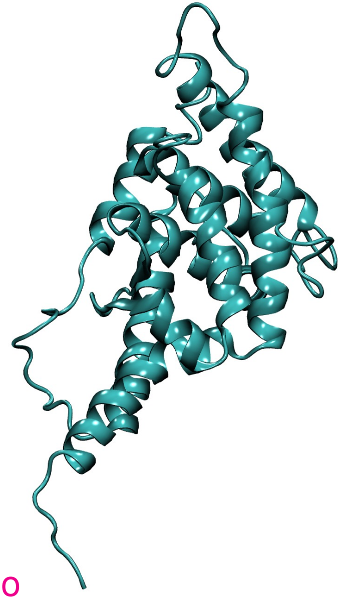
Protein Folding



Protein Folding







Estrogen Receptor: Holo vs Apo

Acknowledgements

Computational Biochemistry Group

Mark Brandt, Chemistry & Biochemistry
Allen Holder, Mathematics
David Goulet, Mathematics
John McSweeny, Mathematics
Elias Eteshola
Abigail Etters
Jacob Hiance
Deborah Lee
Chris Lippelt
Chi Huen Man
Leah Markowitz
Geoffrey Ong
Mitchell Orzech
Jacqueline Simon
Jonathan Taylor

iGEM

Ric Anthony, ABBE
David Goulet, Mathematics
Ben Deschaine
Robert French
Alex Krug
Adam Nighswander
Kristen Schackmann
Devon Trumbauer

Former Students

David Cooper
Kyla Lutz
Melissa Galey
Jonathon Strauser
Vismay Modi