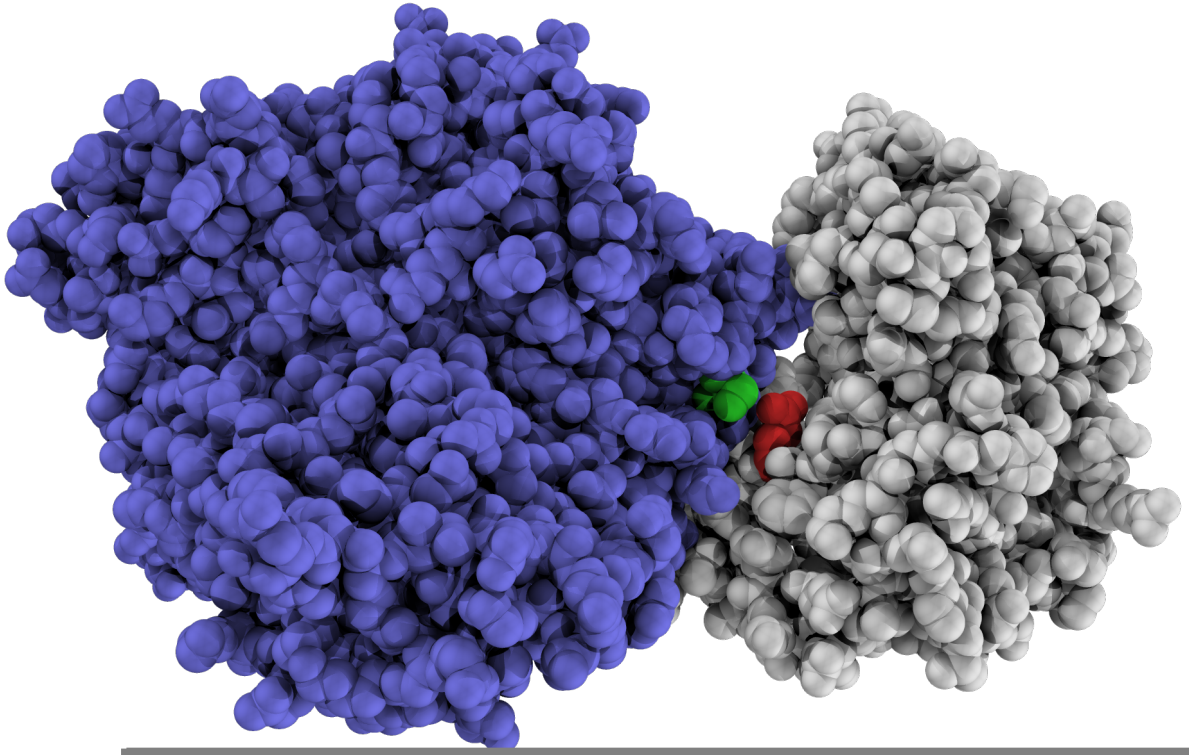
# Introduction:
# Coevolution methods for predicting structure from large numbers of genetic sequences

CS/CME/Biophys/BMI 371

Jan. 18, 2018

Ron Dror

# Amino acids in direct physical contact tend to covary or “coevolve” across related proteins



For example, a mutation that causes one amino acid to get bigger is more likely to preserve protein structure and function (and thus survive) if another amino acid gets smaller to make space

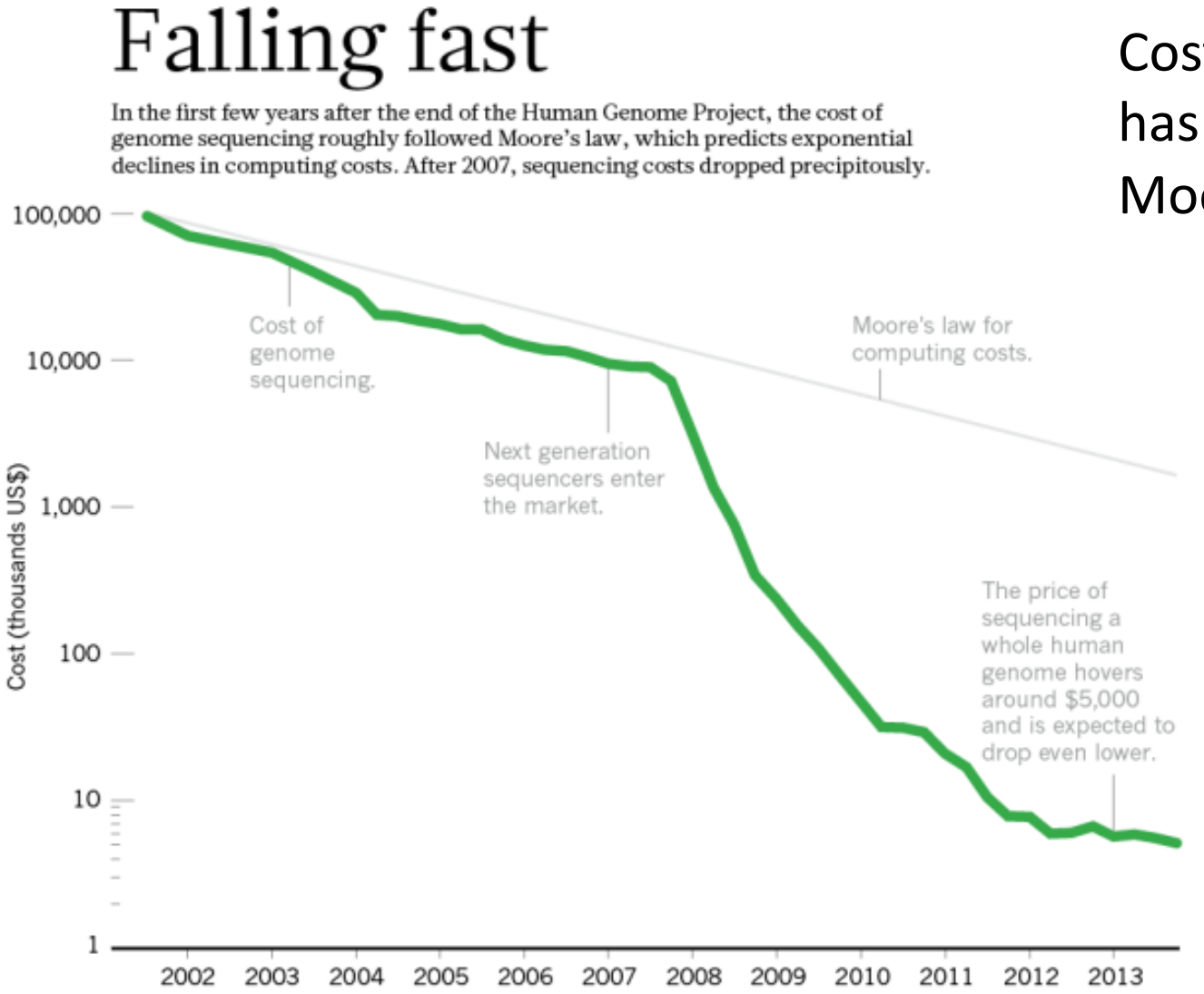
...GANPMHGRDQ**S**GAVASLTSVA...
...GANPMHGRDQ**E**GAVASLTSVA...
...GANPMHGRDE**K**GAVASLTSVG...
...GANPMHGRDS**H**GWLASCLSVA...
...GANPMNGRDV**K**GFVAAGASVA...
...GANPMHGRDR**D**GAVASLTSVA...
...GANPMHGRDQ**V**GAVASLTSVA...
...GANPMHGRDQ**E**GAVASLTSVA...

...VEDLMK**E**VVTYRHFMNASGG...
...VEALMA**R**VLSYRHFMNASGG...
...VATVMK**Q**VMTYRHYLRATGG...
...VARAMR**E**IGKYAQVLKISRG...
...VPELMQ**D**LTSYRHFMNASGG...
...ADHVLR**R**LSDFVPALLPLGG...
...FERART**A**LEAYAAPLRAMGG...
...VPEVMK**K**VMSYRHYLKATGG...

# Can we use this observation to predict structure?

- Given many sequences of related proteins (whose structure is assumed to be similar), look for amino acids that coevolve. They are probably in contact
- This idea has been around for some time, but it started working 6–7 years ago, for two reasons

# Reason 1: Dramatic increase in available sequence data




Cost of genetic sequencing has fallen much faster than Moore's law



E. Hayden, *Nature*, 2014

# Reason 2: Better computational analysis methods

- If amino acid A is in contact with B, and B is in contact with C, then A and C will probably coevolve *even though they are not in contact*
- One shouldn't assume that any two amino acids that coevolve are in contact. Instead one wants a minimal set of contacts that explain the observed coevolution patterns.
- More recent methods exploit additional information.

# Papers for Tuesday

- One of the first papers to demonstrate the utility of these approaches for structure prediction
  - "Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing"
- A recent paper using metagenomics data to determine large numbers of protein structures
  - "Protein Structure determination Using Metagenome Sequence Data"
- A recent paper showing that one can improve performance by using deep learning to incorporate additional information (e.g., the fact that some *patterns* of contacts are more common than others)
  - "Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model"

- For background, see "Protein structure prediction from sequence variation" listed under Additional Papers (https://cs371.stanford.edu/2018_papers/coevolution/additional/nbt.2419.pdf)