Computational Challenges in Genomics and Molecular Biology

Gene Myers

VP, Informatics Research

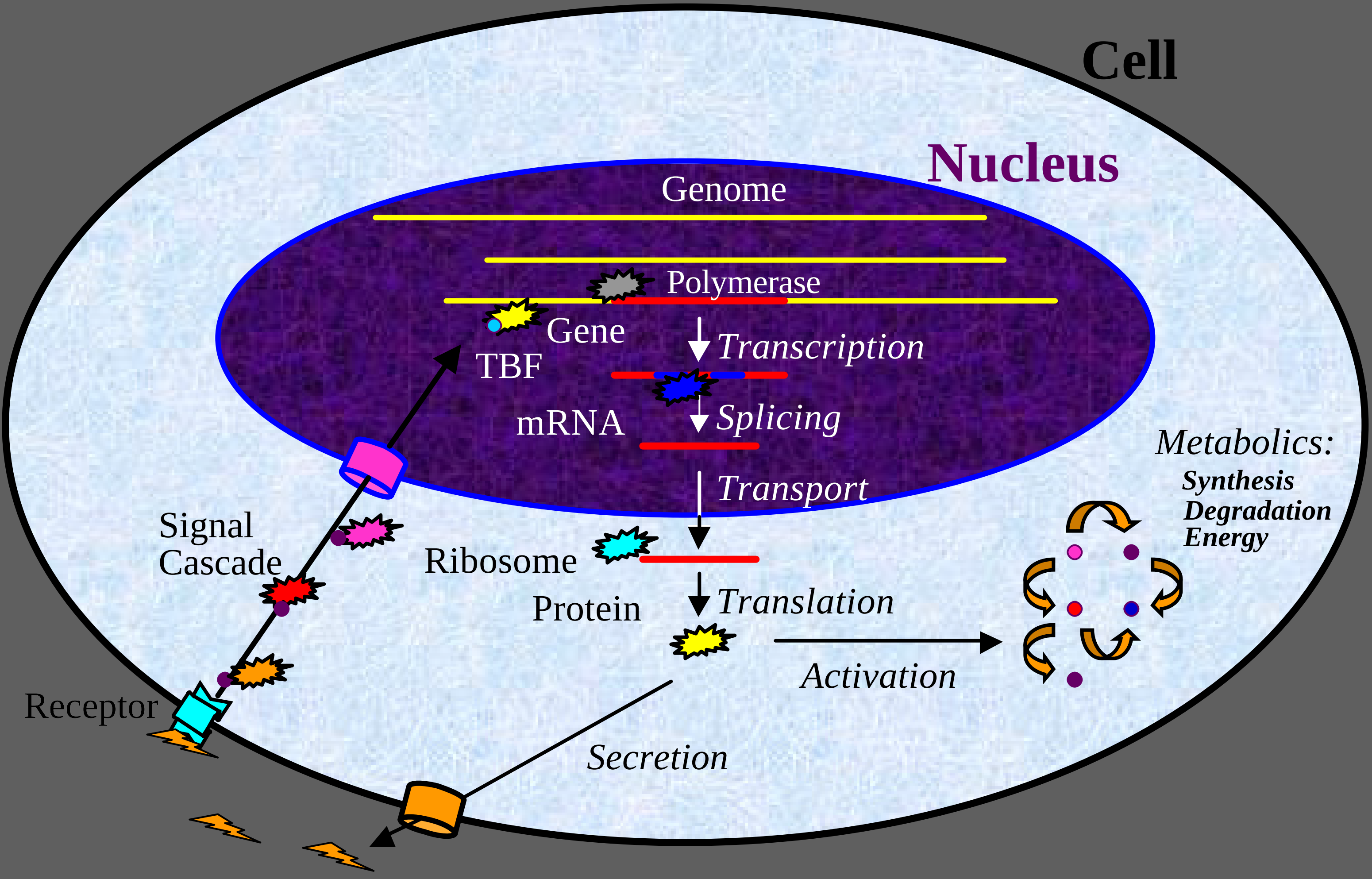
Celera Genomics / Applied Biosystems

The Elements of Molecular Biology

A principal goal is to understand cells and organisms as molecular systems / machines. The basic classes of molecules are:

- DNA in the chromosomes of the genome contains all the information to develop an organism and operate all its cell types.
- RNA serves both short-term informational roles and structural roles.
- Proteins execute the functions of a cell and provides its structural integrity.
- Small metabolites (fats, sugars, etc.) provide energy, raw materials, and serve some limited structural roles.

Cells As Molecular Machines



Understanding Cells at the Molecular Level

- Determining the DNA sequences of the chromosomes of a species.
  Sequencing
- An accurate parts list of all the proteins and RNAs in the cell.
  Annotation
- A graph of all the interactions taking place between these agents.
  Pathways
- What is happening during each interaction.
  Function
- Where each interaction is taking place.
  Subcellular Localization

Current State

- We can sequence the euchromatic portions of genomes.

- We can recognize 75% of the genes but not accurately unless they have been experimentally verified. We don’t know much about alternate splicing.

- We can crudely observe expression of mRNAs and with even greater difficulty observe the more abundant proteins.

- Most accurate molecular biological information is still being verified one hypothesis at a time.

- We must either coordinate efforts or reduce experimental costs to the point where each investigator is greatly empowered.

Current Technologies

- Sequencing: Randomly sample and sequence 600bp stretches from the ends of segments of a given length and assemble, followed by a directed finishing phase.

- Expression Assays: High density arrays where each spot is a set of 18-50bp DNAs complementary to the RNA sequence to be measured, or geometric amplification from a pair of DNA probes complementary to the RNA sequence (quantitative PCR).

- Proteomics: Mass spectrometers can measure the amount and atomic weight of ionized protein pieces (peptides) allowing complex mixtures to be analyzed.

- Light Microscopy: With confocal microscopes and antibody, or RNA, or organo-metallic staining, phenomenon involving but a few particles are being observed.

- All of these technologies involve interesting problems in the interpretation of the data.

  Data Analysis vs. Data Mining

The Role of Informatics

- We need to make computers easier to program – i.e. we need to put scientific computing in the hands of the scientists.
- Our information management technologies are inadequate – huge data sets, semi-structured, data contains errors, not integrated – we need to model these and develop flexible data mining capabilities over them.
- There will be a continued need for new algorithms and tools as driven by new technologies and protocols.
- Physical simulations systems of various types will be needed – docking, ligand binding, stochastic differential equations.
- Experimental design, driven by analysis and simulation, should be a part of our discipline and is an area where we can but are not contributing.

A View of the Future

- Data generation is outpacing Moore’s law by a large margin, but most computations are trivially parallelizable.
- What will you do when a human genome can be sequenced in a couple of hours for $5,000?
- What can you do when protein structures can be routinely determined at modest cost?
- What will you do when nanotech methods exist for probing the cell at the single molecule level?
- The future will be shaped by technology development