Information Extraction in the Biomedical Domain : Biological Principles and IT Resources

Alejandro Pironti
Günter Neumann

News

- The website for the course is ready: http://www.dfki.de/~neumann/bioSeminar2008/
- There is still one fully available seminar topic: Gene Name Normalization at BioCreative Challenge 2

Outline

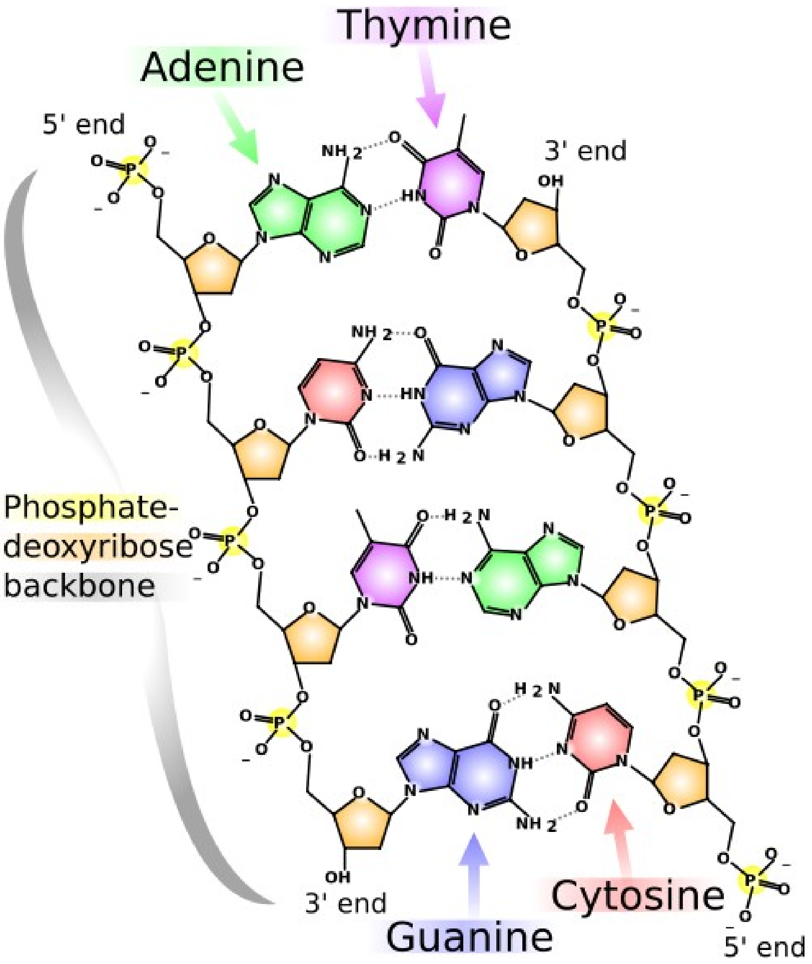
1. Two Experimental Jobs
    - DNA Sequencing
    - Determination of Protein-Protein Interactions

- IT Resources

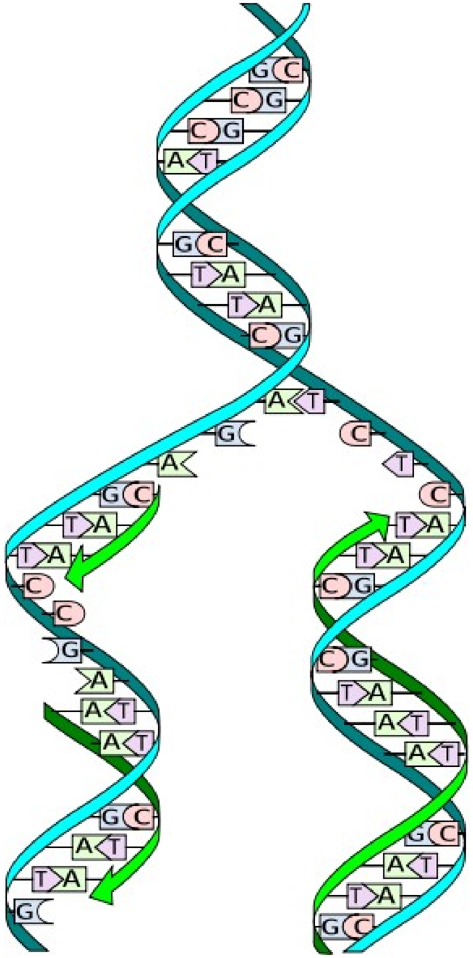
DNA Sequencing

- DNA Sequencing is the determination of the sequence of base pairs in a DNA molecule
- Many applications, e.g. research, diagnostics, forensics
- Several DNA sequencing strategies available:
  - **Chain Termination Methods (Sanger Sequencing)**
  - Pyrosequencing

Review: DNA Structure

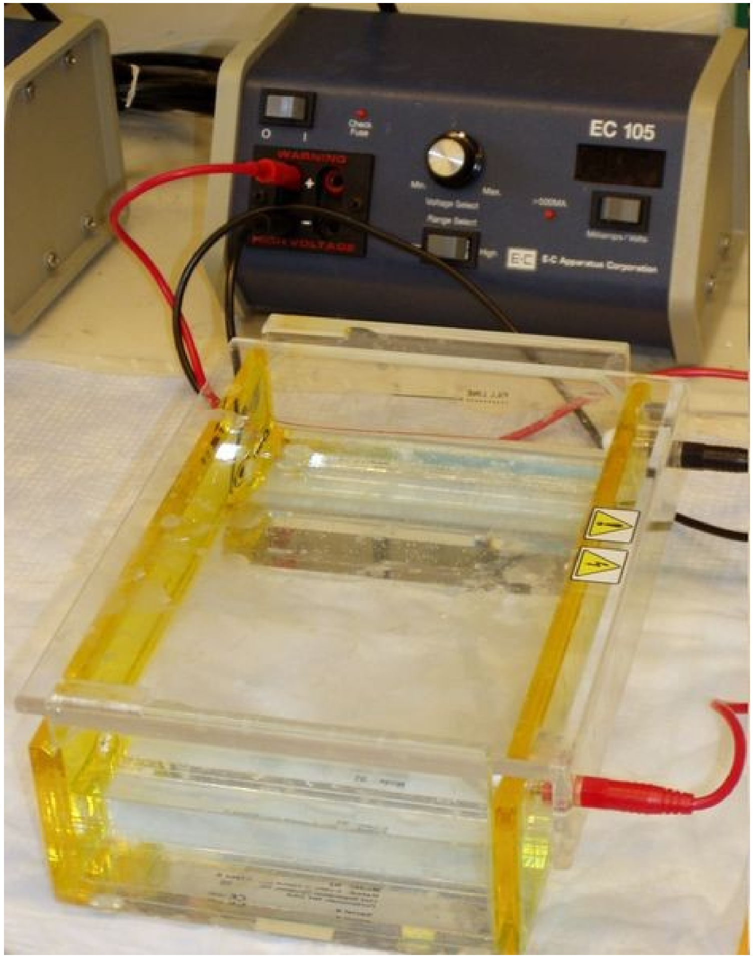


DNA Polymerase Can Replicate DNA



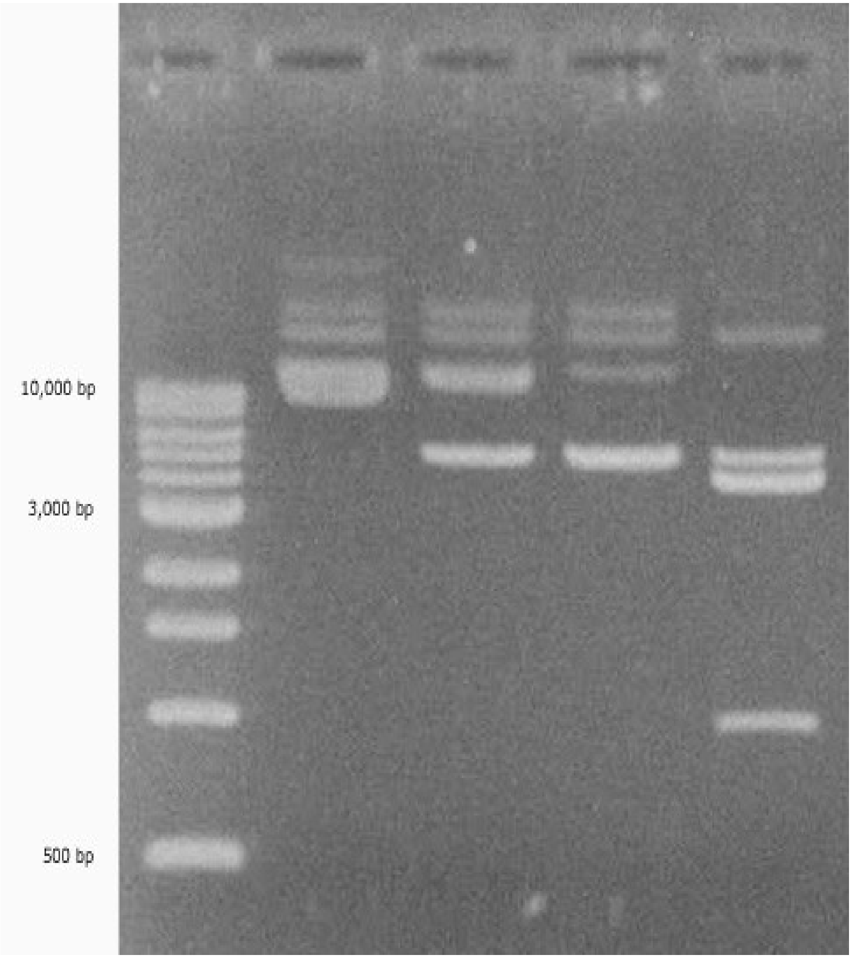
Gel Electrophoresis

- DNA is negatively charged
- This characteristic can be used to separate DNA according to weight in an electric field and an agarose gel
- Small pieces travel faster than big pieces

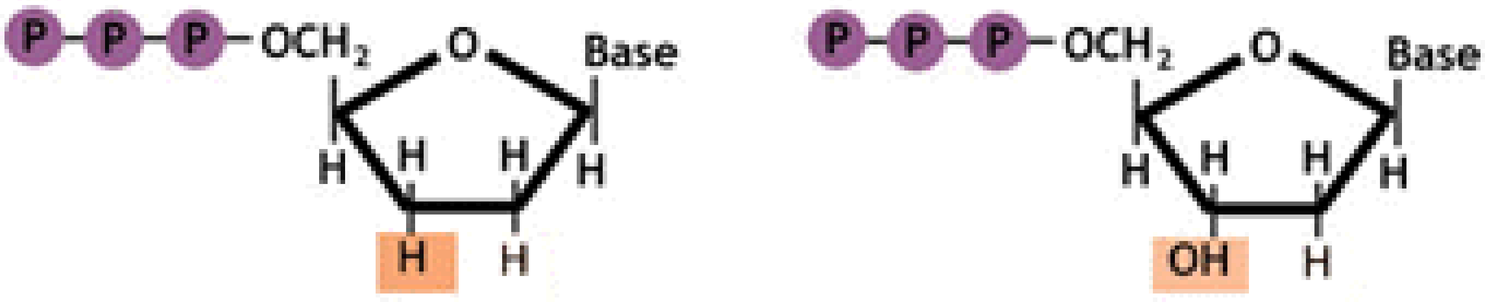


Gel Electrophoresis

- The result of the separation can be made visible with a fluorescent dye (ethidium bromide)



Dideoxynucleotides



dideoxynucleotide (ddNTP)

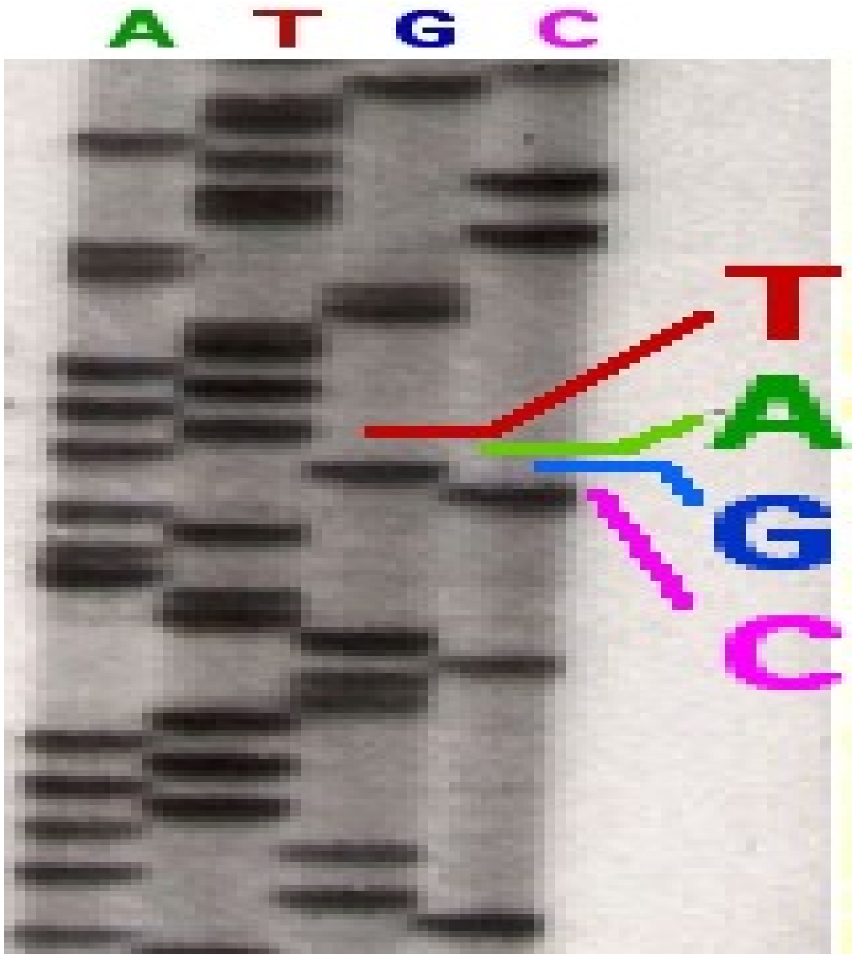
deoxynucleotide (dNTP)

The difference between the stopnucleotide ddNTP and a normal nucleotide dNTP

How Sanger Sequencing Works

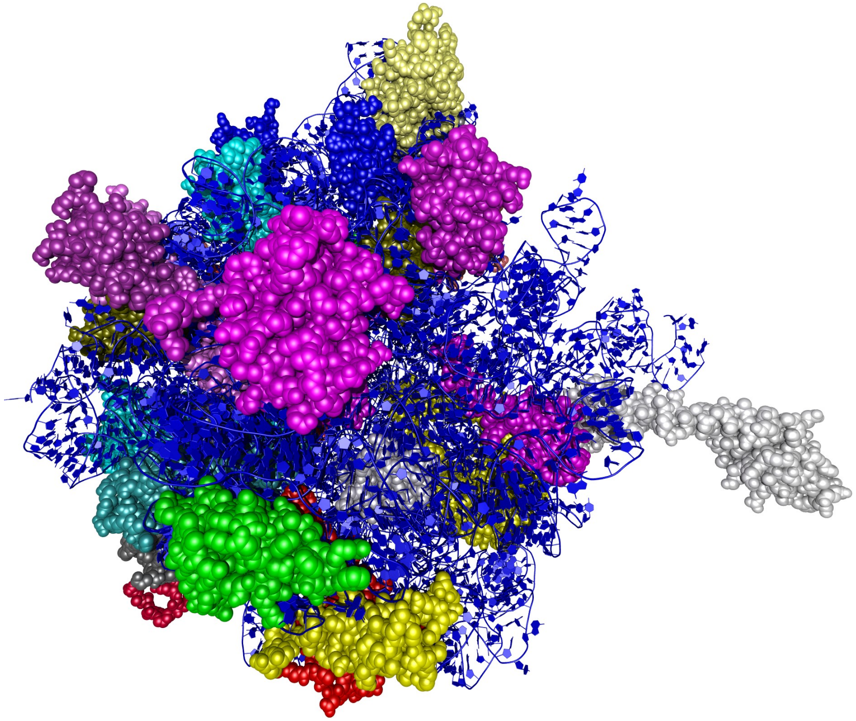
How Sanger Sequencing Works

- Separation of the resulting DNA strands in an agarose gel
- The sequence can be read



Protein-Protein Interactions

- Proteins can interact with each other
- They can form complexes (dock into each other)
- They can assemble to big protein machineries



Methods to Detect Protein-Protein Interactions

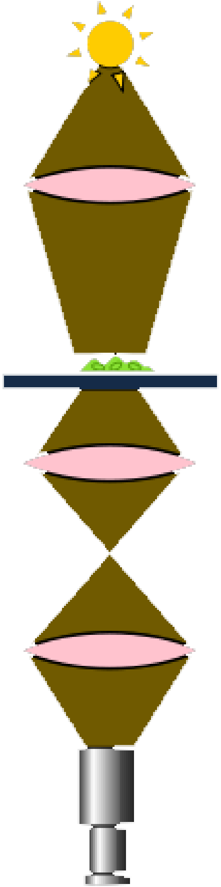
- X-ray crystallography
- Yeast two-hybrid test
- And many other methods!

X-ray crystallography

- Resolution of an elucidated structure: smallest distance between two points such that they can still be recognized as two points
- Resolution depends on wavelength
- The shorter the wavelength, the better the resolution
- Necessary wavelength to resolve atoms: 0.1 nm (X-rays)

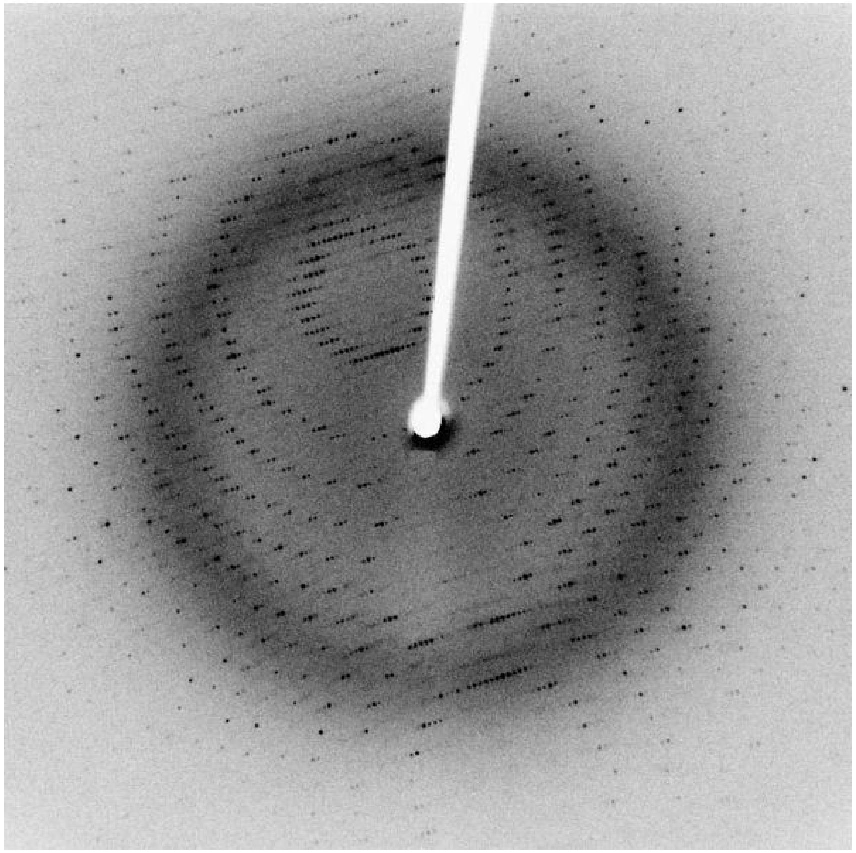
Light Microscope

- There is a light source
- Light is bundled at several points by refractive lenses
- Little problem: there are no refractive lenses for X-rays with such a wavelength!!!!
- And now? Math and physics can help.

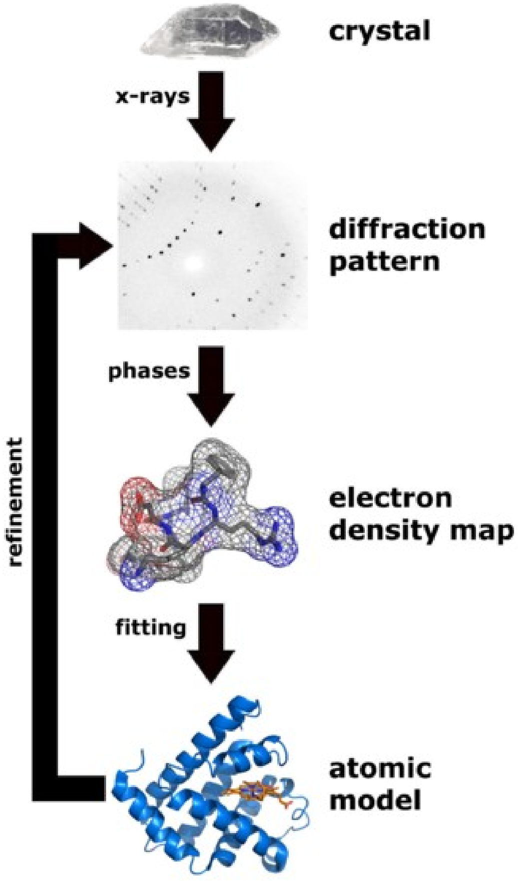


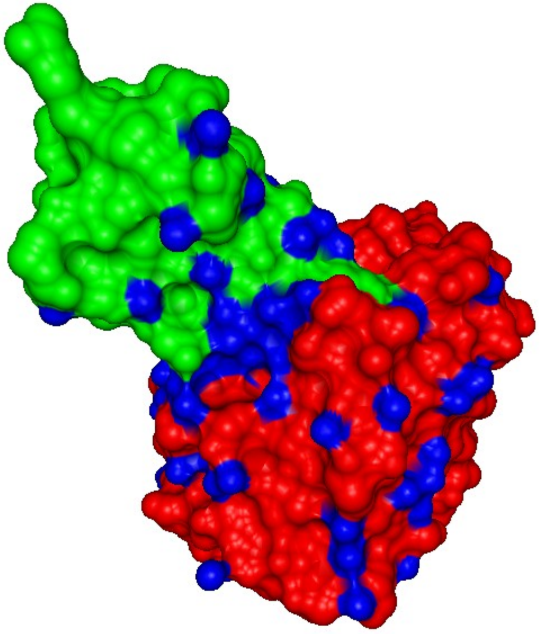
X-ray Crystallography

- Protein crystals: minimize repulsive forces, maximize attractive forces
- Desired effect can be amplified with crystals
- Fire at protein crystals with X-rays
- Register diffraction pattern with photographic film

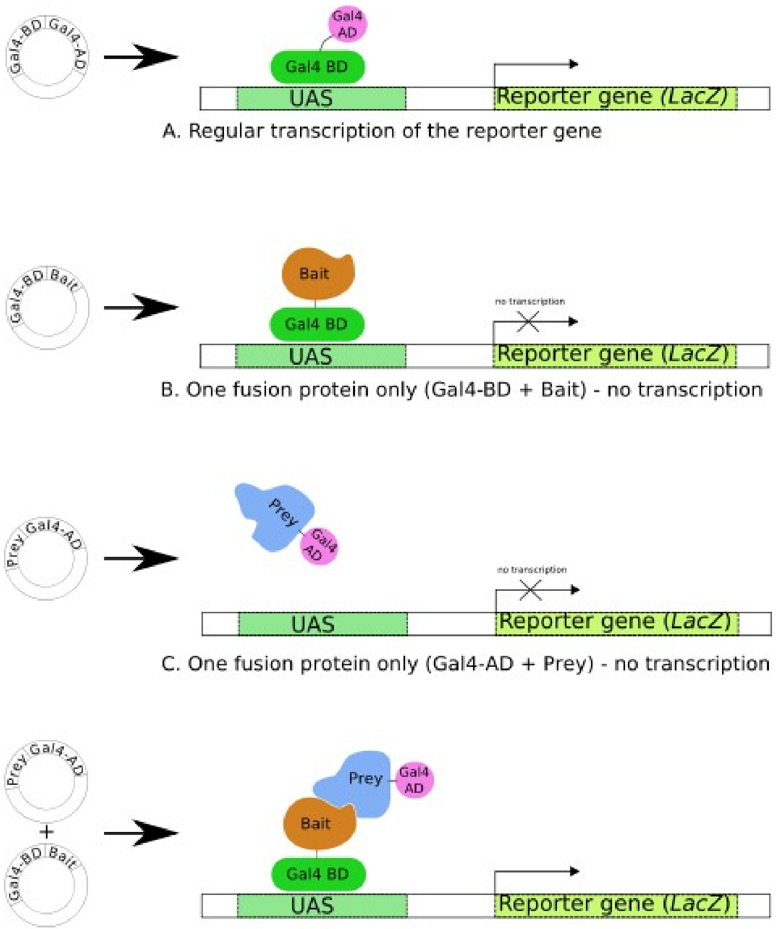


X-ray Crystallography





Yeast Two-Hybrid Test



A. Regular transcription of the reporter gene

B. One fusion protein only (Gal4-BD + Bait) - no transcription

C. One fusion protein only (Gal4-AD + Prey) - no transcription

D. Two fusion proteins with interacting Bait and Prey

2. IT Resources

MEDLINE

- Premier bibliographic database of the National Library for Medicine (U.S.A.)
- Abstracts, Citations
- Can be accessed via PubMed www.pubmed.gov
- Articles indexed my Medical Subject Headings (MeSH)

Controlled Vocabularies

- It is important that annotations are made with exactly the same terms
- Stipulates a default term. Data mining, searches
- E.g. Yeast Two Hybrid, Y2H, Yeast-Two-Hybrid, etc.

MeSH

- Medical Subject Headings
- Controlled vocabulary with a hierarchical structure (various levels of specificity)
- 16 trees describing different areas related to medicine
- http://www.nlm.nih.gov/mesh/meshhome.html