Comparative Immunology

- When did the adaptive immune system evolve?
- Are there several other forms of the adaptive immune system?
- Do the immune system of the plants and mammals have anything in common?
- Such questions can be addressed by studying the evolution of individual genes

Today's computer exercise

How old are NK cells?

How do we tackle this?
1. Blast human E4BP4 to see how many homologs we can find
2. Choose the “good” sequences from the BLAST output
3. We will align all the selected sequences
4. We will make a phylogenetic tree

The basic leucine zipper transcription factor E4BP4 is essential for natural killer cell development

Dorcas M. González1,2, Efrem Long1,3, Henrique Vale-Fernando1,2, Jasper de Boer3, Owen Williams3, Hendrik Sebald1, Mark Colgrove3, Thibaut Krawczuk1, B Hugli1, M Boeck1,3

Natural killer (NK) cells are a subset of lymphocytes crucial for innate immunity and modulation of adaptive immune responses. To contrast to commitment to the T cell or B cell lineage, NK cells have a unique developmental and functional profile. The ability to detect and respond to nonself components in pathogen-infected cells is the key feature of NK cells. Furthermore, NK cells can be activated by antibodies that do not block the antigen or antibody but instead activate the innate immune system.

Specific transcription factors ‘program’ the developmental pathway and the effector function of NK cells. They are divided into different subsets based on their functional and phenotypical properties. The first subset of NK cells, which is induced by IL-12 and IFN-γ, is associated with the expression of high levels of perforin and granzymes. The second subset, which is induced by IL-15 and IL-21, is associated with the expression of high levels of NKG2D and IL-15R. These subsets have different functional properties, such as the ability to induce apoptosis in target cells.

1) Blast: Looking for homolog genes in other species
**Basic Local Alignment Search Tool (BLAST)**

- Heuristic search algorithm
  - Makes shortcuts that are likely (but not guaranteed) to find the optimal hits
- BLAST finds good potential homologs at reasonable speed
  - 10-50x faster than if we align all the sequences
  - More than 100,000 queries per day on the NCBI BLAST server

**Terminology:**
- **Query**: sequence we search the database with
- **Hit** or **Subject**: similar sequence found in the database

**BLAST flavors**

- **Nucleotide-nucleotide searches**
  - Nucleotide database, nucleotide query
  - blastn (start from words of 11 nucleotides)
    - Find homologous genes in different species
  - Megablast (start from words of 28 nucleotides)
    - Designed to efficiently find longer alignments between very similar nucleotide sequences
    - Best tool to find highly identical hits for a query sequence
    - For example: find sequences from the same species
  - Discontiguous Megablast
    - Uses discontiguous words (e.g., W = 11 nucleotides: AT GT AC CG CG T)
    - For example, this can focus the search on codons (the third nucleotide of codons is less conserved due to the degeneracy of the genetic code → next slide)
    - Best tool to find nucleotide-nucleotide hits at larger evolutionary distances for protein-coding query sequences

- **Protein-protein searches**
  - Protein database, protein query sequences
  - blastp (start from words of 3 amino acids)
    - Find homologous proteins in different species
  - Discontiguous Megablast
    - Uses discontiguous words (e.g., W = 11 nucleotides: AT GT AC CG CG T)
    - For example, this can focus the search on codons (the third nucleotide of codons is less conserved due to the degeneracy of the genetic code → next slide)
    - Best tool to find nucleotide-nucleotide hits at larger evolutionary distances for protein-coding query sequences

**The alignment bit-score**

- For a given query, we are mostly interested in finding good hits (highly similar, likely true homologs)
- We could estimate this based on a score derived only from the alignment like the bit-score or percent identity
- … but the chance of finding a hit with a high score by random chance increases if you use a larger database
- … so we have to correct for that

**Expect value (E-value)**

- **E-value**: how many times would you expect a hit this good, by random chance
  - Of course, this depends on the alignment score (S), the length of the query sequence (m), and the size of the database (n): \( E = K m n e^{-\lambda S} \)
  - \( K \): constant for search space scaling
  - \( \lambda \): constant for substitution matrix correction
E-value differs for different databases (because of database size)

Distribution of the number of hits found by random chance in a Large DB

More hits expected by random chance in Large DB

Distribution of the number of hits found by random chance in a Small DB

Fewer hits expected by random chance in Small DB

2) Which sequences should we choose? What is a good E-value?

• This is very difficult to say!
• But as a rule of thumb:
  – E-value <10^-6 for nucleotide blast (blastn, megablast) are good
  – E-value <10^-3 for protein blast (blastp, blastx) are good

• If you want to be very sure that your query and hit sequences are homologs, you should only trust extremely low E-values
• … but sometimes you really have no other information about a protein, except a distant homolog with a very bad (= high, like 0.1) E-value

Today’s computer exercise

How old are NK cells?

How do we tackle this?

1. Blast human E4BP4 to see how many homologs we can find
2. Choose the “good” sequences from the BLAST output
3. We will align all the selected sequences
4. We will make a phylogenetic tree

2) Which sequences should we choose?

• Look for species that you did not yet select, and that you can also say something about the evolution of the NK cells. That is, instead of 10 mammals, choose few birds, reptils, fish, etc
• Check where the HSP, Blast hit, is: is it on a common domain, or on a specific domain. Hits on common domains do not indicate any homology.
Composition of the proteins: Domains

- Proteins often have a modular architecture
  - Consisting of discrete structural and functional regions called domains
- In many cases
  - Different exons code for the different domains in a protein

Today’s computer exercise

How old are NK cells?

How do we tackle this?

1. Blast human E4BP4 to see how many homologs we can find
2. Choose the “good” sequences from the BLAST output
3. We will align all the selected sequences
4. We will make a phylogenetic tree

Which HIV-1 proteins should we use in vaccine?

- ENV?
- Capsid??
Which HIV-1 proteins should we use in vaccine?

- ENV?
- Capsid?

4) Phylogenetic trees

- A phylogenetic tree represents the phylogeny of species or sequences
  - Evolutionary signatures reveal the phylogenetic history
- Phylogenetic trees contain:
  - Present day sequences
  - Ancestral nodes
  - A root
- The same tree can be represented in many different ways:

Speciations and gene duplications

Phylogenetic trees of protein families contain two types of nodes

- Speciation nodes where the protein sequences in the tree diverged due to a speciation event
- Gene duplication nodes where the protein sequences in the tree diverged due to a gene duplication within one genome