1 Introduction

Computational molecular biology is a subfield of computational biology, focusing particularly on the development and application of computational methods for the analysis of data derived more or less directly from the DNA. The field is closely related to another field, *statistical genetics*, which studies statistical modelling and inference in domains typically involving not only data derived from the DNA but also data on various phenotypes and environmental factors; *genetic epidemiology* can regarded as its subfield. Because statistical models easily get quite complex, they often call for computational techniques that can be implemented into computer programs. And vice versa, computational routines rarely suffice for data analysis without a proper statistical model and interpretation. Hence, the fields of computational molecular biology and statistical genetics overlap to a large extent. Consequently, an expert in these areas has to have good know-how of all the three aspects: the biological domain, mathematical and statistical modelling, and computational techniques.

The present course, *Computational Genotype Analysis*, aims to get the student familiar with all these three aspects. The focus is on mathematical/statistical modelling of biological problems in a way that facilitates rigorous mathematical and algorithmic/computational treatment. To this end, we purposefully restrict ourselves to a particular class of biological topics, namely ones that deal with genotype data. On one hand, this means that a great number of important bioinformatics topics—such as gene expression analysis, transcription factor location and gene regulation, metabolism—are not touched at all in this course. On the other hand, the restriction of the scope to a relatively few biological mechanisms and related key notions enables the inclusion of several perspectives on the subject.

2 Practicalities

Some practical issues concerning the course, such as lecture and exercise session hours and the course material, are mostly given on the course's homepage:

http://www.cs.helsinki.fi/courses/582673/2010/s/k/1

Passing the course and grading will be based on both exercises and a home exam, as follows:

Exercises: Three (3) exercises per week given in the lecture notes. One (1) point per exercise that the student "is willing to solve on a whiteboard and the solution is a good try at least." Maximum: 15 points. If the student is not able to attend a particular exrcise session but wants the points, then she/he must send the solutions to the instructor (M.K.) before the session, via email or otherwise.

Home exam: The home exam (or, term paper) will *probably* consist of exercises and a review of an original scientific article related to the courses's topics. Maximum: 15 points. The exam will be announced by Friday Dec 10 and the deadline for the exam will be Friday Dec 17; if these dates are unsuitable to you, please contact the instructor (M.K.) a.s.a.p. for possible rearrangements.

Passing and grading: As follows:

5: 27 points.

4: 24 points.

3: 21 points

2: 18 points.

1: 15 points.

3 Genotypes, haplotypes, and their inheritance

A *genotype* will refer to a genetic material of an individual measured at one or several marker loci. A *marker locus*, or *marker* in short, is a physical location of the genome at which the DNA varies across the individuals in the population; recall that the human DNA is almost identical in every individual. At a marker there are typically two or at most some tens of different variants, called *alleles*, in a population. *SNPs* (single nucleotide polymorphims) are markers of a specific type, locations at which a single nucleotide has undergone a mutation (or, several mutations); typically, and as we will assume throughout the course, a SNP has exactly two alleles.

Diploid organisms, like humans, carry two almost identical copies of the DNA, forming chromosome pairs. Of each chromosome pair, one chromosome of an individual is inherited from the individual's mother and the other from the father. However, it is possible that the chromosome inherited from the mother contains junks of DNA from both parents of the mother, that is, the mother does not necessarily transmit one of her two chromosomes as such to the child but may *recombine* them; the locations of recombination are thought to be rather random and occur independently for every child. An analogous recombination process concerns the DNA inherited from the father.

In many situations it is important to distinguish whether two alleles of a genotype (at different markers) belong to the same member of a chromosome pair, that is, whether the alleles are inherited from the same parent or different parents. To this end, there is a notion of a *haplotype*, which refers to a sequence of alleles along some markers measured from one of the two chromosome in a pair. Thus, a genotype (over some markers) consists of two haplotypes (over the markers). We say that a genotype is *ordered* or *phased* if it explicitly separates the two haplotypes.

For alleles, haplotypes, and genotypes we will mostly use the following notation. Since we focus on SNPs with two alleles, we denote the two alleles usually by A and a, or when multiple loci are considered, then also by B and b, by C and c, and so on. A genotype at a SNP is denoted by AA, Aa, or aa. A typical haplotype over three loci is denoted as AbC, and a haplotype pair or phased genotype as AbC/Abc, the respective (unordered or genotype being $AAbbCc$. Later we will also introduce and use notation based on encoding the two alleles by the numbers (bits) 0 and 1.

See Figure 1 for an illustration of Mendelian inheritance.

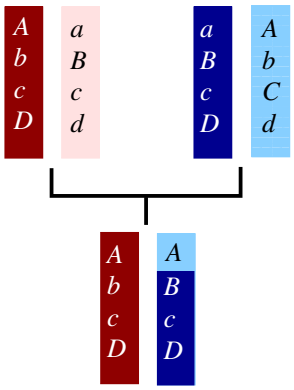

Figure 1: Mendelian inheritance in a mother–father–child trio illustrated at four SNPs. The child inherits the one of the mother's (in red) haplotype as such, whereas the father (in blue) transmits a recombined haplotype to the child.

3.1 Variants of genotype data

All topics considered in the course deal with genotype data that consist of the genotypes of some number n of individuals measured at some number m of markers, usually SNPs. However, the form of the data may vary: first, the individual may be *unrelated*, that is, distant relatives, or they may form mother–father–child *trios*; second, the genotypes may be *phased*, that is, organized into known or estimated haplotype pairs, or *unphased*.

As an example, consider the celebrated HapMap data; see the exercises.

4 Basics of statistical methods

We will use basic probability calculus. For a discrete random variable (vector) X and its possible value x we denote the probability of the event $X = x$ often by $\Pr(x)$. If the probability measure is not fixed but, say, depends on some parameters p, we may write $\Pr(x; p)$.

A *point hypothesis* is indentified with a probability measure, whereas a *composite hypothesis* is identified with a set of probability measures. *Hypothesis testing* aims at deciding whether a given null hypothesis could be rejected based on some observed data x. A common way to implement this is to choose (before looking the actual data) a *test statistic* T, which maps the data X to a single real number $T(X)$. Then the observed data x are associated with a *p-value* $\Pr(T(X) \geq T(x))$; if the p-value is less than some α, then it is said that the hypothesis can be rejected at *significance level* α. Here, for composite hypothesis the probability mesure is usually taken as one of its point hypotheses that maximizes the probability of the observed data x.

5 Basics of algorithms

We will use the following notions rather informally. A computational *problem* is a function from an input space to an output space. An *algorithm* is a sequence of operations that

reads an input and writes an output. An algorithm is said to solve a problem P if for every x in the domain of P the algorithm with input x writes $P(x)$ as the output. The time complexity of an algorithm is the number of basic operations the algorithm executes. The space complexity of an algorithm is the size (usually in bits) of the memory allocated by the algorithm.

Exercises

Take a look at the web page of the International HapMap project,

http://hapmap.ncbi.nlm.nih.gov/index.html.en

and answer the following questions.

I:1 What are the main differences between the different releases?

I:2 Are the data phased or unphased or what?

I:3 How many trios do the data contain?