Context seminar report

Programming Life groep 2

Introduction to molecular biology

Chemical Structure

Deoxyribonucleic acid (DNA) contains **genetic information**. All DNA in an organism is called the **genome**. DNA encodes for all **proteins** needed to live. DNA molecules are linear polymers where each monomer is comprised of a phosphate group, a **nucleotide**, bound by sugar. The monomers only differ in their nucleotide, which is called the **base**. In DNA the sugar is **deoxyribose** and the bases **guanine** (G), **adenosine** (A), **cytosine** (C) and **thymine** (T). In RNA the sugar is **Ribose** and the base T is replaced by **uracil** (U). The structure of the monomers yields a **double-helix form**. Through hydrogen bonding, T and A are paired, G and C are paired. The opposite sides of DNA are therefore **complements**. The non-covalent hydrogen bonds can be broken, resulting into two DNA strands, which is necessary for DNA **replication**. DNA strands can be paired with RNA strands. This is called **hybridization**.

RNA, Transcription and Translation

RNA is usually single stranded and is therefore more flexible, making interactions with itself possible. Some parts of DNA encode for RNA, instead of proteins. A **gene** consists of the region encoding for the protein plus all surrounding control regions. To go from DNA to protein, a polymer, the **genetic information** in one of the two strands of DNA is copied[1] through **transcription** to **messenger RNA** (mRNA); the gene is **expressed**. That strand is called **noncoding strand**[2]. Its complement is the **sense strand**. Genes may overlap, which commonly occurs in **viruses** as a way of packing as much information as possible. RNA encodes for 20 different **amino acids**. Three consecutive bases are called **codons** and encode for one acid. There are multiple encodings for a single acid. Three codons infer a stop signal – or **terminator signal**, which ends of the protein polymer.

Transfer RNA (tRNA) has a three-base **anticodon** and mediates the addition of amino acids to a protein chain. The enzymatic activity that joins amino acids is due to **ribosomal RNA** (rRNA). The process of mRNA - with the aid of tRNA and rRNA - to proteins is called **translation**.

Gene control

Some regions of DNA are **regulatory elements**: control sequences. The control regions where RNA polymerase[3] binds to start transcription are called **promotors**. Other controls are **activators** which improve binding of RNA polymerase and **repressors** which do the opposite. Sequential parts of noncoding DNA are called **introns**, protein-coding sequences are called **exons**. **RNA splicing** removes the introns and mends the exons. **Gene regulation**[4] is mainly the cell types within which genes are activated, their timing and magnitude. These are necessary to regulate **gene expression**. Over- and under expression can have devastating effects. In bacteria gene organization allows for **operons**: sequential encoding for proteins without a stop signal, meaning only one control region for several proteins, that are transcribed to a single piece of mRNA. Operons are rarely found in eukaryotes.

Evolution and mutation

Genes may be changed, added, or destroyed because of **mutations**. Change causes **evolution**. Molecular biologist can transfer individual genes between organisms to produce proteins that some humans are missing because of defects.

[1] From the ‘3 end to the ‘5 end. The mRNA is then translated from the ‘5 to the ‘3 end. Read: only one direction encodes for the correct codons, thus amino acids, thus actual proteins. You can name the ends, so you can infer the transcription and translation directions.

[2] Also antocoding or antisense strand

[3] Enzyme that unbinds DNA so mRNA can be created from DNA

[4] By antisense RNA (aRNA), small interfering RNA (siRNA) and small nuclear RNA (snRNA), the latter also to edit mRNA and maintain chromosome tips (telomeres).

Synthetic gene networks

By combining **naturally occurring genetic components** in unique ways, it has become possible to **artificially** engineer **genetic networks** that possess sophisticated functional capabilities.

Transcriptional control operates at the level of **mRNA synthesis** through the use of **inducible transcriptional activators** and **repressors** that are capable of binding naturally occurring or specifically engineered **promoters**. **Prokaryotic gene control systems** generally use inducible repressors and activators drawn from well-documented **genetic operons** such as the lac operon of Escherichia coli. **Bacterial response regulators** also form the basis of synthetic **eukaryotic** gene regulation systems, although given transcriptional differences they require adaptation.

In considering the **design** of a synthetic genetic network for a **biological application** it is useful to imagine what kind of **functions** one might wish to create. Some applications may benefit from a **mechanism** that ensures a network produces a **consistent** and **stable response**. For other applications, one may require a system that produces **more than one discrete expression state**.

To produce a **unified** and **consistent outcome** a biological process must be **capable of withstanding** a certain degree of **variation** and **difference**. A key development in our understanding of how stability is maintained was through the discovery of **autoregulatory feedback loops** in which proteins, directly or indirectly, influence their own production. An **autofeedback mechanism** can either be **negative**, in which a protein **inhibits** its own production, or **positive**, in which a protein **stimulates** its own production.

The **expression output** of many cell-based regulatory networks is often a **logic** response generated by one or more **input signals**. Their **output** is either **ON** or **OFF** across a wide range of inducer concentrations, except for a small concentration window where transitions between the two states occur. By utilizing several **compatible heterologous gene control systems**, it has been possible to design a range of **eukaryotic logic circuits** that follow strict **Boolean logic** in their integration of two input signals.

To **detect** weak transcriptional responses that, despite being difficult to detect *in vivo*, are often involved in **regulatory functions** where only **trace amounts** of a gene product are required. In typical transcriptional studies aimed at **determining** the **conditions** under which a **promoter** is activated, a **reporter gene** is placed **downstream** of the promoter and assayed under varying conditions. However, where the promoter **response is weak** it is often **not possible** to discern any kind of activity. By placing a **repressor** cascade **downstream** of the promoter it was possible to **amplify** an otherwise undetectable promoter response.

The key requirements for a **band-detection network** (responding to an inducer within a given concentration range) are the design of modular components that enable the detection of a **low-threshold**, a **high threshold**, and a way of **integrating the two thresholds**.

The **pulse-generating network** produces output when a **threshold** concentration is reached, and then through a **feedforward mechanism** shuts down reporter expression regardless of whether the concentration continues to rise or fall. Like the **band-detection network** the **pulse-generating network** provides important insights into how pulse-generating behavior could occur in **natural systems**.

Mathematics of chemical reactions

The **equilibrium position** of a reversible reaction, the point at which the rate of formation equals the rate of dissociation, is determined by a combination of the forward and reverse **rate constants**.

For unidirectional reactions ($A + B \rightarrow X + Y$): If X is the number of molecules of any reactant, the **rate of disappearance** equals that of any other reactant and is decribed by $\frac{dX}{dt}$. Due to the *Conservation of Mass* principle, this rate must be inversely proportional to the rate of appearance of any and every individual product. To calculate the amount of molecules of a species at any time t, the rate of (dis)appearance can be integrated over the interval $[0, t]$ for any reactant or product.

$\frac{dX}{dt} = \frac{dY}{dt} = -\frac{dA}{dt} = -\frac{dB}{dt}$ and $X(t) - X_0 = Y(t) - Y_0 = -A(t) + A_0 = -B(t) - B_0$

To know any one of the amounts is to know all, but to know any one of them, the speed of the reaction must be known. According to the *Law of Mass Action* the **rate of simultaneous combination** (second derivative of X as a function of t) of two chemicals is proportional to the product of their concentration $X]$ and can be described as $k[A][B]$, where k is the **constant of proportionality**.

Assuming the concentration of a species at any moment is either determined by the fixed volume of the medium (closed reaction vessel) or by the inflowing reactants (open reaction vessel), the concentration may either refer to the relative number of molucules of a species within a solution when the species dissolves in the medium, or the ratio of precipitated molecules (only product) to the medium. This notion of concentration at any moment of species X is $x(t)$.

Combining the equations of amount and concentration, and the *Law of Mass Action*, the differential equation with initial value $x(0) = x_0$ is found, which stationary points are given by setting the right-hand side to zero and solving: $\frac{dx}{dt} = kab = k(a + x_0 - x)(b + x_0 - x)$ and $x = a_0 + x_0$ or $x = b_0 + x_0$. By solving the general equation, the **progression of the reaction as a function of time** is found.

For reversible reactions: The principles are the same, but there is a forward constant k_1 and a backward constant k_{-1} for each chemical. If the forward and backward reactions are independent, the rate of change is the sum of the effects of each reaction. If the forward and reverse constants are distinct, the equation describing the stationary points is quadratic and thus easily factorized, with real roots if $x_0 \leq y_0$. The progression of the reaction can be found in a similar way as that of a unidirectional reaction.

Enzyme Kinetics

The reaction from substrate S in combination with enzyme E to product P is performed by forming an **enzyme-substrate complex** C which decomposes into product and enzyme. Due to the typically small amount of enzyme compared to subtrate, the conversion rate is limited when the enzyme becomes saturated with substrate (**enzyme saturation**). No enzyme is destroyed or produced, so e_0 is the initial and total amount of enzyme. The combination $k_M = \frac{k_{-1}+k_2}{k_1}$ of rate constants is known as the **Michaelis-Menten constant** and it determines the concentration of complex c at low substrate concentrations. At high substrate levels, the complex reaches a relatively invariant concentration c_{Eff}, and because $c = \frac{se_0}{k_M+s}$ becomes dependent on s, $c_{Eff} \approx e0$.

The velocity v of the reaction is the appearance rate of the product and so at high substrate levels $v_{max} = k_2e_0$. The initial velocity is found by substituting v_{max} into the *Law of Mass action* for $t =$. By expirimentally measuring the reaction rate for various substrate concentrations a sketch of the graph can be made and working from the graph k_M can be determined, being the substrate concentration where the reaction rate is half maximal.

Synthetic biology and minimal cell

The field of synthetic biology has three enablers that helped it emerge and develop rapidly: **computational modelling**, **DNA sequencing** and **DNA synthesis**.

Synthetic biology, not unlike systems biology, relies heavily on computational modelling. It helps in the **design** and **prediction** of a system prior to fabrication. Synthetic biology can therefore be considered to be the **application** of certain systems biology techniques. However, synthetic biology lacks the software the like are available to a systems biologist, such as an integrated development environment (IDE) to aid in designing synthetic biological systems.

The **quantitative measurement** of biological parameters obtained through systems biology is an essential to accurately design synthetic biological devices and systems. On the other hand, synthetic biology can prove especially useful in **substantiating hypotheses** developed by systems biology, as these hypotheses can be tested in a much more controlled system.

DNA sequencing, or the reading of DNA, is the second enabler for synthetic biology. DNA comprises of four bases: A, T, C and G. Sequencing an entire organism's genome provides researchers with a wealth of **data**. Sequencing can also be used to **verify** that engineered sections of DNA have been **fabricated correctly**.

DNA synthesis is the process of writing, or synthesising, DNA. This is the third key enabler of synthetic biology. There is commercial activity around strands of DNA of about 100 to 1000 base pairs, either **constructed** by **combining bioparts** or by **synthesising the pairs directly**. The typical cost of fabricating strands of DNA have dropped significantly over the years, and are currently around 55 cents per base pair. At present there is a **technological barrier** around the cost and speed of synthesising, which would have to be removed to make the design process loop in synthetic biology commercially viable.

Another challenge is **synthesising chromosomes of mammals**. In bacteria the DNA strands are circular (plasmids) and consist of about tens of thousands of base pairs. In mammals, DNA lies in packed in chromosomes which each contain millions of base pairs. These chromosomes are much harder to produce than plasmids.

The **yields** of synthesised DNA sequences decrease dramatically the longer the sequence is. The longer a strand, the more prone it is to sequencing errors. The theoretical yields of a 20, 100, 200 and 1000 base pair strand of DNA respectively are 90%, 50%, 35% and 5%. To increase the yields the industry requires the development of new techniques, which is ongoing. Apart from computational modelling, DNA sequencing and DNA synthesis a number of additional tools are used which are important to the development of synthetic biology. A **chassis** is a host cell for the synthesized DNA. It's important the chassis be **genetically simple**. One of the most used natural chassis is the Escherichia coli or E. coli bacterium. It can be grown easily and has relatively simple genetics. Other common chassis are B. subtilis, Mycoplasma, Yeast and P. putida.

An alternative approach to natural chassis is to create **minical cells**. The basic idea behind minimal cells is to produce a cell which has the minimum number of components required to support biological synthesis from synthetic DNA circuits or genomes, whilst being as simple as possible in order to achieve adequate control of its function.

A key part in synthetic biology is determining the **minimal genome** for sustaining cell life. Synthesizing a minimal cell is key for many hypothesis in replication and its subsystems. The model is to be created in a tube, *in vitro*. Because a minimal cell is **highly complex**, some of the subsystems would be substituted. If

the model turns out to be a poor model for the more complex *in vivo* system, the model can be made more complex.

To **determine the minimal set of genes** for a self-replicating system we can try to search for genes that have homologs in the genomes of groups of organisms. A challenge here is that over long evolutionary distances genes have replaced each other and a third of the essential genes have unknown functions. **Genetics** search for essential genes by mutating one gene at the time. Some genes will be falsely marked essential because the mutation mutilates neighboring genes or creates a new complex structure. Finally there is **biochemistry** which identifies from cell fraction those gene products essential for the reconstruction of biochemical reactions. So far this yielded the best results.

By **reconstituting** the catalysts that synthesize DNA, RNA and protein, an MCP may be realized. Since **lipids** can be used to act as simple membranes, this draft is a major milestone. Another method entirely is a **cell free approach**. In this strategy, only biochemical extracts containing the components necessary to operate the synthetic DNA circuit are employed. This method offers more control of the synthetic device as the normal processes of the cell won't interfere with the intended processes.

What is Bioinformatics?

Bioinformatics is conceptualising biology in terms of molecules and applying **informatics techniques** to **understand** and **organise** the **information** associated with these molecules, on a **large scale**. There are **three** aims of bioinformatics: First, bioinformatics organises **data** in a way that allows researchers to access existing information and to submit new entries as they are produced. The second aim is to develop **tools** and **resources** that aid in the analysis of said data. The third aim is to use these tools to **analyse** the data and **interpret** the results in a biologically meaningful manner.

Much of the data found can be grouped together based on biologically meaningful similarities. For example, distinct proteins frequently have comparable **sequences**. There are common terms to describe the relationship between pairs of proteins or the genes from which they are derived: **analogous** proteins have related folds, but unrelated sequences, while **homologous** proteins are both sequentially and structurally similar. Homologues can be distinguished between **orthologues**, proteins in different species that have evolved from a common ancestral gene, and **paralogues**, proteins that are related by gene duplication within a genome.

From these observations, a finite '**parts list**' can be made for different organisms: an inventory of proteins contained within an organism, arranged according to different properties. With this list, categorising the proteins, for example by folds, results in a simplification of the contents of a genome, useful for future genomic analyses.

Protein sequence databases are categorised as **primary, composite** or **secondary**. Primary databases function as a repository for the raw data. Composite databases compile and filter sequence data from different primary database to produce combined non-redundant sets. Secondary databases contain information derived from protein sequences and help the user determine whether a new sequence belongs to a known protein family.

A source of **genomic-scale data** has been from **expression experiments**, which quantify the expression levels of individual genes. These experiments measure the amount of mRNA or protein products that are produced by the cell.

The most profitable research in bioinformatics often results from integrating multiple sources of data. However, it is not always easy to access and cross-reference these sources of data because of differences in naming and file formats.

In the end, two principal approaches form the basis of all studies in bioinformatics **(the three aims)**: **comparing** and **grouping** the data according to **biologically meaningful similarities** and **analysing** one type of data to infer and understand the observations for another type of data, so we can **understand** and **organise** the **information** associated with biological molecules on a **large scale**.

Bioinformatics Challenges for Personalized Medicine

Single Nucleotide Polymorphisms (SNPs) are now recognized as the main cause of human genetic variability. By combining these genetic associations with phenotypes and drug response, **personalized medicine** will tailor treatments to the patients' specific genotype. In the coming years, the bioinformatics world will be inundated with **individual genomic data**. This flood of data introduces significant challenges that the bioinformatics community needs to address and which fall in the following four main areas.

1: Processing large-scale robust genomic data

Sequencing technologies are becoming affordable and are replacing the microarray based genotyping methods. The error rate from these technologies is a source of significant challenges in applications, **including discovering novel variants**. A remaining challenge for short read assemblers is **reference sequence bias**: reads that more closely resemble the reference sequence are more likely to successfully map as compared with reads that contain valid mismatches. When the **diploid sequence** is known, reference sequence bias can be

avoided by mapping the reads to both strands. Another challenge is **developing new methods** for novel SNP discovery. Finally, there is a pressing need to **improve quality control** metrics.

2: Interpreting the functional effect and the impact of genomic variation

After genomic data has been processed, the **functional effect and the impact** of the genetic variations must be analyzed. In the last few years, several **computational methods** have been developed to predict deleterious missense SNPs. **Prediction methods** do not provide any information about the pathophysiology of the diseases and so experimental tests are required to validate genetic predictions. The methods for the **analysis** of SNPs are mainly limited to the prediction of the impact of missense SNPs.

3: Integrating systems data to relate complex genetic interactions with phenotypes

Given the **complex phenotypes** involved in personalized medicine, the simple "one-SNP, one-phenotype" approach taken by most studies is insufficient. Given the size of the genomic data sets, **dimensionality reduction methods** will be essential to make complexity algorithms tractable. **Systems biology** and **network approaches** address to the problem of complexity by **integrating molecular data** at multiple levels of biology. Combining disparate data sources can result in novel associations and provide insight into gene-gene and gene-environment interactions.

4: Translating these discoveries into medical practice

The ultimate challenge for this research is to apply the results for improved patient care. Most pharmaceutical development addresses medical problems with a "one drug fits all" approach. **Pharmacogenomics** connects genotype to patient specific treatment and has already been successful in improving drug prescription and dosing. Bioinformatics also translates discoveries to the clinic by **disseminating discoveries** through curated, searchable databases. Ultimately, bioinformatics needs to develop methods that **interrogate the genome** in the clinic and allow physicians to use personalized medicine in their daily practice.

Trends in computational biology

Next-generation sequence analysis

Computational biology has risen in prominence in recent years largely because of the increase in the data-generation capacity of high-throughput technology.

The advance: Two methods for de novo transcriptome assembly of short reads were published.

What it means: Advances in transcript assembly from RNA-Seq data should allow alternative splicing to be studied genome-wide across many biological conditions. In addition, RNA-Seq is now poised as a tool for discovering new RNAs that we may not have even known were transcribed from a genome.

Discovery from data repositories

Electronic medical records are becoming a reality, promising lower health-care costs, improved patient treatments and, perhaps, scientific advances.

The advance: A first study that demonstrates the feasibility of associating genetic modifications with data on phenotypic traits mined from electronic medical records.

What it means: The case of electronic medical records illustrates the potential value locked within unique biomedical databases and the challenges of realizing that value.

Learning to see

In biological research, one advantage of computational analysis is automation and fidelity.

The advance: Machine-learning algorithms that could accurately classify whether a pattern of fluorescent staining represents localization to one subcellular organelle or to a mixture of locations.

What it means: This represents the first step toward a new way of thinking about interpreting images that is generative rather than descriptive. Whereas descriptive approach may take an image of a cell and tell you that the protein is in the nucleus, a generative approach builds a model for other images.