Electr Eng (2007) 90:161–173
DOI 10.1007/s00202-007-0062-6

ORIGINAL PAPER

Information and communication theory in molecular biology

Pavol Hanus · Bernhard Goebel · Janis Dingel · Johanna Weindl · Juergen Zech · Zaher Dawy · Joachim Hagenauer · Jakob C. Mueller

Received: 5 January 2007 / Accepted: 13 January 2007 / Published online: 4 April 2007

Abstract The DNA sequencing efforts of the past years together with rapid progress in sequencing technology have generated a huge amount of sequence data available in public molecular databases. This recent development makes it statistically feasible to apply universal concepts from Shannon's information theory to problems in molecular biology, e.g to use mutual information for gene mapping and phylogenetic classification. Additionally, the genetic information in the cell is continuously subject to mutations. However, it has to be passed from generation to generation with high fidelity, raising the question of existence of error protection and correction mechanisms similar to those used in technical communication systems. Finally, better understanding of genetic information processing on the molecular level in the cell can be acquired by looking for parallels to well established models in communication theory, e.g. there exist analogies between gene expression and frame synchronization.

P. Hanus (✉) · B. Goebel · J. Dingel · J. Weindl · J. Hagenauer
Institute for Communications Engineering, TU-München,
Arcisstr. 21, 80290 Munich, Germany
e-mail: Pavol.Hanus@tum.de

J. Zech
Institute for Statistical Medicine and Epidemiology,
TU-München, Clinic "Rechts der Isar", Ismaninger Str. 22,
81675 Munich, Germany

Z. Dawy
Department of Electrical and Computer Engineering,
American University of Beirut, P.O. Box 11-0236,
Riad El-Solh, Beirut 1107 2020, Lebanon

J. C. Mueller
Max Planck Institute for Ornithology, Postfach 1564,
Haus Nr. 5, 82319 Seewiesen, Germany

Keywords Information theory · Communication theory · Molecular genetics · Classification · Gene mapping · Error correction · Frame synchronization

1 Introduction

Communications engineering as well as genetics have both experienced a major breakthrough in the mid twentieth century. In 1953, the double helix structure of the DNA was deciphered by Watson and Crick. From this point on it was clear that the genetic information is stored in form of two complementary directed strands composed of letters from a four symbol alphabet. Until the discovery of the molecular basis of genetics, the research was concentrating on classical genetics, based on the rules of Mendelian inheritance of traits. Shannon [21] himself was using mathematics to study how different trait combinations propagated through several generations of breeding in his Ph.D thesis completed in 1940. He devised a general expression for the distribution of several linked traits in a population after multiple generations under a random mating system, which was original at that time, but went largely unnoticed, since he did not publish his work. After completing his Ph.D thesis, Shannon shifted his focus towards digital communications and cryptography.

In 1948, Shannon [22] established the theoretical fundamentals of digital communication systems. He introduced the concept of information based solely on the statistical characteristics of the information source. He defined information in an abstract way independent of semantics that does not differentiate between text, video or audio as was generally being done when studying

communication systems at that time. Using such information definition, Shannon proved that a message generated by an information source can be losslessly compressed to the entropy of the source (source coding theorem) and that it is possible to code the information in a way, such that one can transmit it error-free at the maximum rate that the channel allows (channel coding theorem). Ever since, communications engineers have been devising algorithms to achieve the limits of these two theorems. The definition of information based solely on statistical characteristics of the information source also applies to genetic data. Recent advances in DNA sequencing technology supply enough data to apply Shannon's general information concept to molecular biology. Section 2 gives a short introduction to basic principles from molecular biology required for better understanding of the following sections. In Sect. 3 we show how mutual information and compression can be used for phylogenetic classification. Section 4 describes the application of mutual information to gene mapping. The question whether an error correcting code has evolved on the genome sequence level is addressed in Sect. 5. Finally, in Sect. 6 we model transcription initiation (one step in protein synthesis) as frame synchronization in a communication system.

The original involvement of information theorists with molecular genetics goes back to the discovery of the genetic code. In the period between the discovery of the DNA structure in 1953 and the decipherment of the genetic code 1961–1969, when no actual DNA sequences and only very few amino acid sequences were known, several different coding schemes describing the mapping of the DNA sequence (four letter alphabet) to a protein (amino acid sequence from a 20 letter alphabet) were proposed by coding theory experts. Some of them had high information density, while others have foreseen error correction capabilities. The experimental discovery of the actual genetic code (the mapping rule of the $4^3 = 64$ DNA sequence triplets to the 20 amino acids and a stop symbol) was a disappointment for the coding community since it does not seem to implement any of the two. A review of the proposed codes can be found in [12]. From this point, there has been little interaction between the two communities until recently. We believe that with all the newly available sequence data further interactions could be fruitful as our research suggests. The question why the genetic code has evolved the way it is remains open. There seems to be evidence for the optimality of the code in terms of error minimization using metrics based on physio-chemical properties of the resulting amino acids like their hydrophobicity [10]. Apparently, evolution imposes additional constraints on the optimization of how the genetic information is being stored, which makes the modeling rather peculiar. This has to be accounted for by communications engineers modeling evolution and the molecular processing of genetic information in the cell as a communication system.

2 Biological background

2.1 DNA

In 1944, the desoxyribonucleic acid was identified as the primary carrier of genetic information. The discovery of the geometric arrangement of the DNA building blocks in a double helix by Watson and Crick followed in 1953. The DNA consists of two complementary directed strands of nucleotides. Each nucleotide is composed of a backbone unit (sugar and phosphate) and one of the four bases adenine (A), guanine (G), cytosine (C) or thymine (T). The sugar phosphate backbone determines the direction of each strand which is referred to as 5′ to 3′ by convention. The two strands are held together by electrostatic interaction via weak hydrogen bonds between the complementary bases A–T and C–G, see DNA in Fig. 1. Here, nature has implemented a simple complementary repetition code, which is very advantageous for DNA replication, that has to take place every time a cell divides. Each of the two complementary strands is used as template for the DNA copy of one of the two daughter cells.

2.2 Mutations

The process of copying is prone to errors leading to point-mutations, insertions, deletions and duplications. According to evolutionary theory a certain degree of mutation is necessary to allow for adaptation of different species to changing environmental conditions. Propagation of evolutionary disadvantageous mutations is hindered by natural selection in contrast to neutral and the rare advantageous mutations. Assuming a common ancestor, the degree of dissimilarity in the genomes of existing species can be used to reconstruct their phylogenetic relationships, as shown in Sect. 3. Mutational variations observed across the human population are the origin of genetically influenced diseases. The main objective of gene mapping is to determine which of the varying positions in the genome, also referred to as single nucleotide polymorphisms (SNPs) [1] are related to the disease under investigation. Section 4 describes an information theoretical method to identify the SNPs which are statistically related to the investigated disease. It relies on population based data from clinical stud-

Fig. 1 Protein synthesis

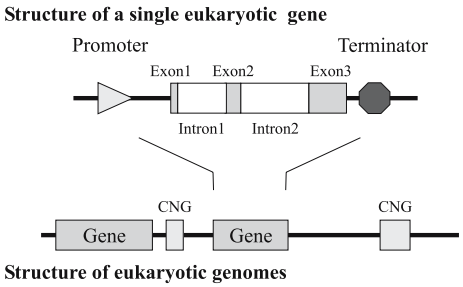

Fig. 2 Genome organization of multicellular organisms

ies. Since high rate of mutation would lead to too many evolutionary disadvantageous mutations per generation cycle, it is crucial that the genome copying process takes place with high fidelity. Nature has implemented mechanisms to minimize the error susceptibility of the copying machinery. However, error protecting measures on the sequence level similar to error correcting codes in communication systems are currently not known. We believe that especially in case of complex multicellular (eukaryotic) organisms, which have long generation cycles and a limited number of offsprings, nature might have developed sequence level error correcting measures to ensure the necessary high replication fidelity. The primary and best understood function of the genome is to carry information for the synthesis of proteins, see Sect. 2.3. However, in complex eukaryotes like vertebrate the proportion of the genome actually coding for proteins is less than 10%, as opposed to simple fast evolving single cell organisms (prokaryotes), where almost all of the genome codes for proteins. The non-coding part has been largely neglected by the research community for a long time until comparative genomics has recently identified regions in the genomes of vertebrate species that do not code for proteins, but show a high degree of evolutionary conservation [26], labeled conserved non-genic region (CNG) in Fig. 2. This implies some unknown evolutionary important function. The proportion of such conserved non-coding regions in the human genome is comparable to that of protein coding regions. Currently, our search for error protecting means on the sequence level concentrates on these regions, see Sect. 5. They might be carrying parity information to protect the coding regions.

2.3 Protein synthesis

The protein coding part of the genome is converted to proteins in a process called gene expression. It takes place in two basic steps, see Fig. 1. First, during transcription the genomic DNA region coding for a protein is copied into messenger RNA (mRNA) by the RNA polymerase molecule. The resulting mRNA corresponds to a complementary copy of the template strand except that the base `T` (thymine) is substituted by `U` (uracil). In the second step, the ribosome molecule translates the mRNA into a sequence of amino acids—a protein. Hereby, triplets of bases are converted to amino acids according to the mapping rule described by the genetic code [19].

2.4 Genome structure

The protein coding portion of the genome is arranged in genes. The genes vary in size and are randomly distributed across the genome. The beginning of a gene is characterized by a promoter sequence in front of it. The end is signalled by a terminator. During transcription initiation, the first step in protein synthesis, the promoter sequence has to be detected. This resembles frame synchronization in digital communication systems. Further investigation of this analogy is presented in Sect. 6. In eukaryotes the mRNA produced during transcription contains non-coding regions called introns. These are being spliced out (removed from the mRNA) before translation occurs. Only the coding exons are finally translated to protein. The described genome structure is depicted in Fig. 2. The content recognition method described in Sect. 3 can be used to distinguish between the coding exons, non-coding but transcribed introns and the non-genic regions not taking part in gene expression.

3 DNA classification using compression distance measures based on mutual information

The possibility of using mutual information for classification and content recognition of genetic sequences is exploited in this section. Two different mutual information based distance measures are proposed, one for classification and one for content recognition. The measure

proposed for classification is a metric. The influence of compression based entropy estimation on the proposed measures is investigated. Examples of successful applications in the field of genetics are presented.

Mutual information describes the amount of information shared by stochastic processes. It can be used to derive distance measures quantifying the similarity of the processes. Mutual information based distance measures can be used to compare texts written by different authors or to build phylogenies of different species.

3.1 Compression based entropy approximation

The definition of mutual information is based on the entropies of the compared sources, which will be approximated using compression. The idea of using compression for phylogenetic classification of whole genomes was first introduced in [14]. Shannon's fundamental theorem on data compression states that every source S can be losslessly compressed up to its entropy rate $H(S)$. Thus, the compression ratio achieved by an optimal compression algorithm designed for a given source S when compressing a message s generated by this source is a good approximation of the sources actual entropy rate

$$H(S) \approx \frac{|\mathrm{comp}(s)|}{|s|}, \tag{1}$$

where $|.|$ denotes the size in bits or symbols. The entropy of DNA sequences is less than two bit due to the use of a four symbol alphabet (`A,C,G,T`).

In general a universal compressor for a whole class of sources (e.g. DNA sequences, natural texts) is available. Such universal compressors gradually adjust their underlying general statistical model describing the whole class of sources to the individual statistics of the particular message being compressed. For example, genomic DNA sources contain approximate repeats and palindromes (reverse complements) due to duplications and point mutations that occur during evolution. DNA-compress uses this general property of genomic DNA and compresses the specific repeats occurring in the particular sequence being compressed. Such universal compressors are particularly suited to compare sources of a given class as they should be able to compress well a concatenation of messages generated by similar sources as opposed to dissimilar ones. Consequently, the conditional entropy $H(S_i|S_j)$ of two different sources S_i and S_j will be approximated as the compression ratio achieved for the message s_i when the compressor's model is trained on the message s_j. The compression size of the concatenated sequences $|\mathrm{comp}(s_j, s_i)|$ can be used for this purpose

$$H(S_i|S_j) \approx \frac{|\mathrm{comp}(s_j, s_i)| - |\mathrm{comp}(s_j)|}{|s_i|}. \tag{2}$$

3.2 Mutual information based distance measures

The aim of unsupervised classification is to build clusters of all sources S_i based on chosen criteria. A distance metric $d(S_i, S_j)$ quantifying the similarity of the sources is required for such clustering.

Content recognition serves a different purpose. Here, a set C of known content sources $S_i^C, i \in \{1 \ldots |C|\}$ is provided together with a set U of unknown sources $S_j^U, j \in \{1 \ldots |U|\}$. The goal is to find the best matching content source S_b^C with the smallest distance $b = \arg\min_i(d(S_i^C, S_j^U))$ for each unknown source S_j^U. The distance measure for content recognition on the contrary to classification does not have to satisfy the axioms of a metric.

Information theory describes the relatedness of sources S_i and S_j as the mutual information $I(S_i; S_j)$ shared by these sources

$$I(S_i; S_j) = H(S_i) - H(S_i|S_j) = I(S_j; S_i). \tag{3}$$

Mutual information is an absolute measure of information common to both sources. It can be transformed to a bounded distance through normalization in two different ways: one way, to be used for content recognition, is to normalize by the maximum possible mutual information the two sources can share, resulting in

$$d_{CR}(S_i, S_j) = 1 - \frac{I(S_i; S_j)}{\min(H(S_i), H(S_j))} \leq 1. \tag{4}$$

The lower bound is reached for sources that share the maximum possible mutual information given their entropies. It can be reformulated using conditional entropies

$$d_{CR}(S_i, S_j) = \frac{\min(H(S_i|S_j), H(S_j|S_i))}{\min(H(S_i), H(S_j))}. \tag{5}$$

Using the compression based approximations in (1) and (2) it can be written as

$$d_{CR} = \frac{|\mathrm{comp}(s_j, s_i)| - |\mathrm{comp}(s_j)|}{|\mathrm{comp}(s_i)|}, \tag{6}$$

for $|\mathrm{comp}(s_i)| < |\mathrm{comp}(s_j)|$. Since the triangle inequality is not satisfied for d_{CR} this measure is not a metric distance. Thus for classification we normalize $I(S_i; S_j)$ by the maximum entropy of both sources resulting in the following distance metric

$$d_{CL}(S_i, S_j) = 1 - \frac{I(S_i; S_j)}{\max(H(S_i), H(S_j))} \leq 1. \tag{7}$$

Compared to d_{CR} in (4) the two sources must not only share maximum possible mutual information, but also need to have identical entropies in order to achieve $d_{CL} = 0$.

The advantage of the compression based approximation of the derived distances is that no prior alignment of the compared sequences s_i and s_j is necessary.

3.3 Results

Different types of compression algorithms were tested with respect to their classification and content recognition performance: Lempel–Ziv, Context Tree Weighting, Burrows Wheeler Transform, Prediction by Partial Matching (PPM) and DNACompress. In general PPM and DNACompress performed best for genetic sequences. A set of properties making a compression algorithm suitable for classification and content recognition was derived in [7].

A typical classification problem in molecular genetics is reconstruction of phylogenetic relationships between different populations (e.g. human populations, different mammalian species) in form of a binary tree, where the nodes represent the separation events and the root the common ancestor of all the investigated populations according to the evolutionary theory. Figure 3 shows a phylogenetic tree of the human population constructed using d_{CL} with DNACompress and the quartet tree generation method described in [4]. Mitochondrial DNA (mtDNA) was used for this study. It is about 16,000 bases long and particularly suited for phylogenetic studies, since it is inherited only maternally and shows high rate of mutation because it resides in mitochondria outside of the cells protecting nucleus. The migration pattern observed in the tree corresponds to the currently accepted theory of African human origin and the results presented in [27]. Interesting highlight is the close relationship between North American Navaho descendants and the European Finnish population, indicating that North America might have not only been populated from north eastern Asia by crossing the Bering land bridge, but possibly also through the Arctic.

To demonstrate the content recognition performance of the derived measure, we present the results for content recognition of non-genic regions (ng), exons (ex) and introns (in). As content sequences the first 50,000 nucleotides (50 kb) of concatenated sequences of each type were taken from the human chromosome 19 (c19). Sequences of different sizes of each type taken from the beginning of chromosome 1 (c1) were used as unknown sequences. For each unknown sequence j the distance $d_{CR}(S_i^C, S_j^U)$ to every content sequence i was calculated. Using DNACompress and d_{CR} all unknown sequences

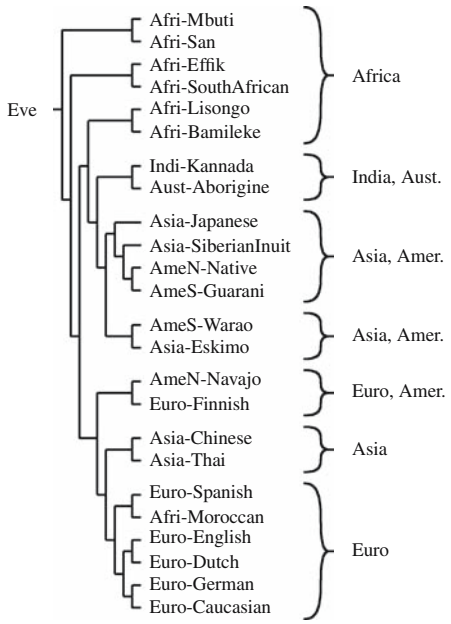

Fig. 3 Human phylogeny based on mtDNA

Table 1 Content recognition of non-genic regions (*ng*), introns (in) and exons (ex)

$S_j^U \setminus S_i^C$	c19ng-50kb	c19in-50kb	c19ex-50kb
c1ng-300kb	0.04-best	0.84	1.02
c1ng-13kb	0.65-best	1.01	1.01
c1in-300kb	0.93	0.58-best	1.01
c1in-13kb	1.00	0.05-best	1.07
c1ex-300kb	1.02	1.01	0.96-best
c1ex-13kb	0.98	0.94	0.83-best

were recognized correctly as shown in Table 1. Some distances are greater than 1 due to the concatenation in the compression based approximation of conditional entropy in (2), leading to high compression ratios if a dissimilar sequence is used for training.

The obtained results demonstrate how the derived distance measures approximated using compression can successfully be applied to phylogenetics and recognition of sequence type. In Sect. 4 the d_{CL} distance measure will be used for pairwise SNP comparison in gene mapping.

4 Gene mapping and marker clustering using Shannon's mutual information

This section discusses the application of Shannon's information theory to population-based gene mapping. In addition, a mutual information based distance measure is used in conjunction with multidimensional scaling to build and visualize clusters of genetic markers. The

presented approaches are applied to clinical data on autoimmune Graves' disease.

Mutual information, defined as

$$I(X;Y) = \sum_x \sum_y p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}, \tag{8}$$

where X and Y are random variables, can be interpreted as the reduction in entropy (or uncertainty) of one random variable given another. In the following, it will be used as a measure of dependence between the physical manifestation of a trait (phenotype) and the underlying genetic make-up (genotype). Connecting particular phenotypes with the causal genotypes is the main aim of gene mapping.

4.1 Gene mapping

About 90% of deviations between the genomes of two individuals from a population are single point mutations. Such variations in the genomes of a population occurring with a relative frequency $\geq$1% are referred to as single nucleotide polymorphisms (SNPs). It is estimated that only about 0.3% of the human genome are SNPs. The term allele refers to the nucleotide observed at a particular SNP locus (position) in an individual. At most one mutation per genome position is assumed to have occurred during the short human evolution. This assumption results in biallelic SNPs—exactly two different alleles are observable per SNP in a population. Assume that a particular region of the genome was sequenced across the population resulting in ACCGTA in 76% of the cases and ATCGTA in 24%. The second position would thus be a SNP with major allele A = C and minor allele a = T. In a simplistic view sexually reproducing organisms posses two homologous copies of their genome, each inherited from one of the parents. Thus, per SNP locus we observe two alleles, one from each parental side. An individual will have either inherited two homozygous alleles from both parents (either AA or aa) or two heterozygous (a different allele from each parent Aa or aA). Modeled as a discrete random variable each SNP locus would thus have four possible realizations. However, the genotyping does not allow to distinguish the parental origin of the alleles (Aa is indistinguishable from aA), reducing the number of observable realizations of a SNP to three.

In a typical clinical population-based gene-mapping study a small subset of L suspect SNP markers from the overall estimated 10 million human SNP loci $S_1, S_2, \ldots, S_L$ is genotyped in N individuals. Preferably, in a population-based disease study half of the individuals (the cases) carry the disease under investigation, the other half (the controls) are healthy. In such case-control studies the phenotype P is a binary variable (healthy/diseased) and the genotype a set of ternary random variables $S_1, S_2, \ldots, S_L$. Figure 4 depicts a simple channel diagram describing the information transfer from a ternary SNP S_i to the binary phenotype P. The probabilities of the random variables' realizations and the transition probabilities can be derived from relative frequencies, i.e. observed counts divided by N. These probability estimates exhibit a variance that depends on the sample size N. From these probabilities, the mutual information $I(S_i;P)$, where $i = 1 \ldots L$ between each SNP S_i and the phenotype P can be estimated to investigate each SNPs causality both in absolute (through the unit bits) and relative terms [18].

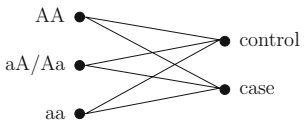

Fig. 4 Genotype–phenotype transition diagram for a two-locus model

The approach presented here for single SNPs and binary phenotypes can be easily extended to the joint analysis of multiple SNPs and/or higher order and continuous phenotypes [6]. A detailed comparison of the proposed method with other statistical and signal processing based methods can be found in [20].

4.2 Marker clustering

So far, we have used mutual information between phenotype and genotype. In this section, we use the mutual information between SNPs to find groups or clusters of correlated genetic markers, which are likely to form evolutionary entities. This is an important tool for gene mapping, as it can provide additional hints about which markers should be interpreted jointly. As distance measure between two SNPs the metric $d_{CL}(S_i, S_j)$ presented in (7) is applied. In this case S_i and S_j represent SNPs. In order to avoid biased results, only the data from the controls should be used to compute the distances between all markers. Subsequently, classical multidimensional scaling can be used to cluster and visualize the SNPs in two- or three-dimensional space for further analysis [5].

4.3 Results

The proposed methods were successfully tested on simulated and real data sets. The clinical data set described in [25] was used to generate the results presented in this section. The study suspects a 317 kb long region across

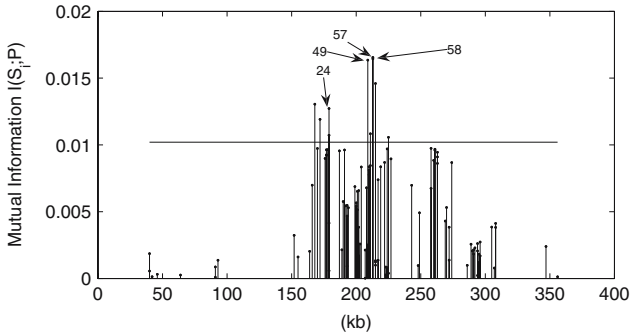

Fig. 5 Mutual information in *bit* btw. Graves' disease and each SNP in a region suspect of being related to the autoimmune disease

the genes CD28, CTLA4 and ICOS to be related to the Graves' autoimmune disease. The region comprises 108 dispersed SNP loci, which were genotyped in 384 cases and 652 controls. Figure 5 shows the mutual information estimate for all 108 SNPs [6]. It should be noted that the effects measured are relatively weak (≈0.01 bit as compared to the theoretical maximum of 1 bit). To determine the results' significance, the permutation-based critical values of the total study (global null hypothesis based on 5% significance level) have been determined and plotted. Our analysis of the autoimmune disease data set also revealed two study-wise significantly associated regions, which are identical to the most promising regions found by the logistic regression analyses reported in [25]. The multidimensional scaling clustering analysis (described in Sect. 4.2) of the same dataset in two-dimensional space is depicted in Fig. 6. It can be seen that the SNP loci identified as significantly related to the Graves' disease tend to cluster. The resulting cluster indicated by the ellipse in Fig. 6 points to similar evolutionary histories and ages of these markers. This implies with high probability only a single causal marker among these SNP loci.

In comparison to other statistical gene-mapping methods, applying the simple, yet theoretically well-defined concept of mutual information to the representation of SNP-phenotype and SNP–SNP relationships does not require any assumptions to be made and thus lays out a consistent framework for a first screen in gene mapping approaches.

5 Conserved non-genic elements—implementations of error correcting codes?

The DNA is the primary carrier of genetic information. This information must be "transmitted" to various destinations. During cell replication the genomic information must be copied and passed on to the two daughter cells as each cell carries a copy of the whole genome. A further example is the transmission of genetic information from genes to proteins. The genetic transmission channels introduce noise and one might ask whether nature has developed error protecting means similar to those that we use in digital data transmission over noisy channels in order to make reliable communication possible? Consider the transmission of genetic information over generations in evolutionary time. The DNA is subjected to mutations making this transmission channel noisy. Assuming a simple model of nucleotide mutations, Battail [3] showed that the capacity of this channel decreases exponentially over time. He concludes that, for any reasonable instantaneous mutation rate, genome conservation over large geological timescales can only

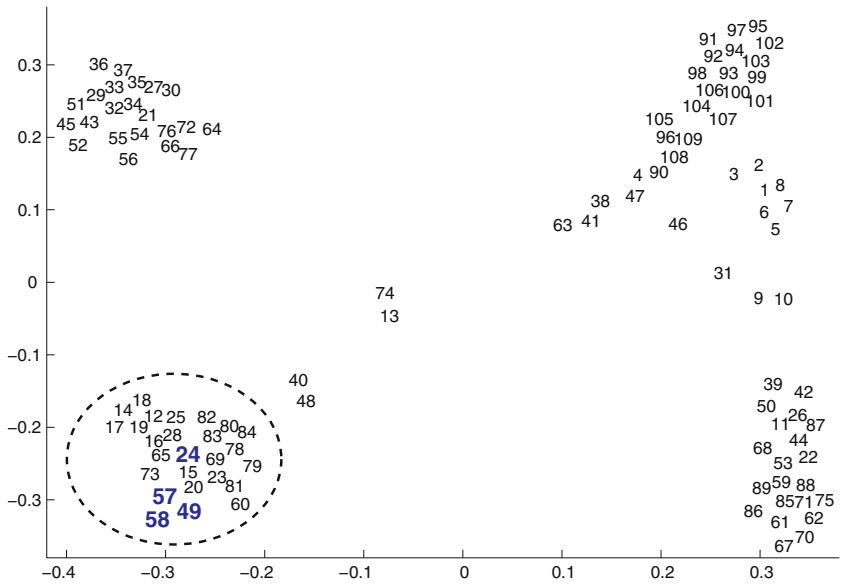

Fig. 6 Marker clustering in a two-dimensional scaling; autoimmune disease data

be explained by genome regeneration. He hypothesizes that there exists an error correcting code implemented on the genome sequence level for this purpose and that genome regeneration must occur before the capacity of the channel falls below the error correction ability of this code [3]. Further evidence for this hypothesis is given by a recently published discovery about the error correcting ability of the plant *Arabidopsis* [15]. The experiment shows that mutations that are present in the genomes of the parents are corrected in the genomes of their offsprings with certain probability. Assuming an error correcting code on the genome level, we need to find out where in the genome it is implemented. This amounts to the detection of functional elements in the genome, i.e. separating evolutionary noise from meaningful biological information.

5.1 Detection of functional DNA sequences and conserved non-genic elements

Comprehensive identification of biologically functional elements in the DNA represents a central and ambitious goal in modern genetics. The reliable detection and analysis of functional elements are crucial steps towards a deep understanding of how complex organisms work. Early approaches to this problem were limited to the use of information from one species. Today, with high quality genome sequences of several species at hand, a comparative approach, taking into account multiple sources of information, is often used to infer regions in the genome subjected to evolutionary pressure. The evolutionary relationship of multiple organisms can be described in form of a phylogenetic tree. The common ancestor is represented by the root of the tree. The passage of DNA along the organismal lineages is described by the branches of the tree. During the process of evolution, the passed genetic information (DNA) is subjected to mutations that cause variations. Natural selection decides about the success of the transmitted DNA. Altered information in regions whose variation will negatively influence the fitness of the organism will most likely diminish the organisms capability to reproduce and prevent passing its DNA to the next generation whereas mutations in regions not being under selective pressure will be passed on to further generations without restrictions. Thus, those elements within the genome carrying information for important basic functions are less likely to successfully mutate during evolution due to natural selection. Consequently, by identifying conserved elements in the assembly of the genomes of several species, we find candidates that are very likely to be functional. Nowadays, having access to the complete sequences of a number of vertebrate genomes this approach provides a powerful tool for the systematic discovery of functional elements in the genome [9,17,24].

5.2 Evolution in a communication theoretic framework

In terms of communications engineering the evolution can be regarded as a single input multiple output system. In the biological transmission system (evolution), we may think of the common ancestor as the transmitter. Its sequence of bases is the output of the information source. In Fig. 7, a single input multiple output communication scenario and an evolution scenario are depicted. The divergence of lineages, indicated by the inner nodes of the phylogenetic tree, is equivalent to the scattering of the dispersing electromagnetic wave on obstacles. The leaves of the tree correspond to the receiver antennas in the SIMO system. They receive the sequences that we are able to observe in the species today. The information is transmitted over the branches of the phylogenetic tree, equivalent to the signal paths in terms of communications theory. Errors (mutations), erasures and insertions occur during transmission.

5.3 Modeling evolution

Commonly, the evolution can be described by a set of parameters [28]. We abstract evolution by a phylogenetic tree $\mathcal{T} = \{\boldsymbol{\tau}, \boldsymbol{t}'\}$ that we specify by a topology $\boldsymbol{\tau}$ and the respective branch lengths $\boldsymbol{t}'$ accounting for the phylogenetic relationships and the evolutionary distances among the species. A continuous time stationary Markov process with state space $\mathcal{X} = \{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$ describes the mutation process. A rate matrix $\boldsymbol{R}$ defines this Markov process and is related to the matrix of transition probabilities between two nodes in the phylogenetic tree by

$$\boldsymbol{P}(t_{u\to v}) = e^{\boldsymbol{R}t_{u\to v}}, \tag{9}$$

where $t_{u\to v}$ denotes the evolutionary distance between the nodes u and v in the tree. The rate of substitutions at a site is strongly dependent on its position along the DNA sequence as some regions are under purifying selection

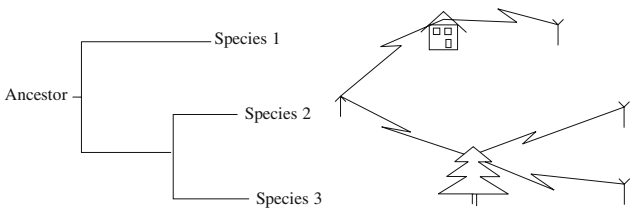

Fig. 7 *Left* Phylogenetic tree relating three species as they evolved from a common ancestor. *Right* A single input multiple output scenario

and thus evolve more slowly than neutral regions. In terms of the phylogenetic description, we model rate heterogeneity as a site dependent scaling parameter θ_i, where i denotes the nucleotide position, working on the lengths of the branches of the tree.

$$\boldsymbol{t}_i = \theta_i \boldsymbol{t}'.$$

The thus influenced absolute evolutionary distances lead to higher or smaller substitution probabilities according to (9) and more or less conserved regions. In the following, we parameterize evolution by the set $\boldsymbol{\psi}_i$ containing the parameters described above

$$\boldsymbol{\psi}_i = \{\boldsymbol{R}, \boldsymbol{\tau}, \theta_i \boldsymbol{t}'\}. \tag{10}$$

Note that evolution is site dependent, theoretically each site i could evolve differently. However, in practice over large regions of the genome constant values for $\boldsymbol{R}$ and $\mathcal{T}$ are assumed.

5.4 Estimation algorithm

Figure 8 shows the transmission model for evolution. The single sequence $\{x_i\}$ is transmitted over the multipath channel evolution. At the receiver, we observe the receive vector sequence $\{\boldsymbol{y}_i\}$ consisting of the ancestral sequence as we observe it today in the genomes of the considered species. The channel is characterized by the transition probabilities $p_{\boldsymbol{y}}(\boldsymbol{y}_i|x_i;\boldsymbol{\psi}_i)$ conditional on x_i and parameterized over $\boldsymbol{\psi}_i$. The channel is not constant for all input sequences. Different genome regions have been subjected to different substitution rates because they are subjected to different natural selection pressure dependant of the biological importance of the information they carry. From this point of view, estimating the conservation of a particular DNA region is equivalent to the estimation of how good the transmission channel was in this region. We will introduce a detection method which, in contrast to earlier approaches [17,24], is independent of the assumption about neutral evolutionary rates and which does not require a priori tuning parameters. We propose a definition of conservation that relies on the Kullback–Leibler distance to the well defined maximum possible conservation that does not allow for any mutations to occur [11]. From a communication theoretic viewpoint, the maximum conservation is equivalent to the case of noiseless transmission, i.e the base x_i is observed unchanged in all components of the receive vector $\boldsymbol{y}_i$. In this situation, the channel shall be specified by $p_{\boldsymbol{y}}(\boldsymbol{y}_i|x_i;\boldsymbol{\psi}^0)$ and the receive vector $\boldsymbol{y}_i$ is distributed according to $p_{\boldsymbol{y}}(\boldsymbol{y}_i;\boldsymbol{\psi}^0)$. For the comparison with the maximum conservation case, we estimate the evolutionary model that maximizes the likelihood of an ensemble of received vectors. In a sliding window over the observed data $\boldsymbol{Y}_i = [\boldsymbol{y}_{i-\delta},\ldots,\boldsymbol{y}_{i+\delta}]$, δ fixed, we determine the evolutionary model $\hat{\boldsymbol{\psi}}_i$ that most likely led to the observed data. Assuming statistical independence among the columns of $\boldsymbol{Y}_i$

$$\hat{\boldsymbol{\psi}}_i = \arg\max_{\boldsymbol{\psi}_i} \left\{ \sum_{j=i-\delta}^{i+\delta} \log(p_{\boldsymbol{y}}(\boldsymbol{y}_j;\boldsymbol{\psi}_i)) \right\}. \tag{11}$$

We calculate the probability mass function $p_{\boldsymbol{y}}(\boldsymbol{y}_i;\hat{\boldsymbol{\psi}}_i)$ for a column parameterized by $\hat{\boldsymbol{\psi}}_i$ and compare the estimated distribution with the one corresponding to the maximum conservation process using the Kullback–Leibler distance

$$s_i = \mathcal{D}\left(p_{\boldsymbol{y}}(\boldsymbol{y}_i;\hat{\boldsymbol{\psi}}_i)||p_{\boldsymbol{y}}(\boldsymbol{y}_i;\boldsymbol{\psi}^0)\right). \tag{12}$$

s_i is the score assigned to the column in the middle of the sliding window. Note that a low score corresponds to a good channel and thus a highly conserved region. A score of zero is best explained (in the ML sense) by the process of maximum conservation. Gaps are treated as missing data causing the algorithm to consider only the subtree of species where data is available. A comparison of the results that we obtained with our method is presented in the next Section. Figure 9 shows our estimation of conservation and the underlying genomic data, and alignment of the genomes of five species. Mutations are highlighted by colored background. Our distance based score signal reflects the different degrees of conservation as one can observe by comparing the signal course with the data. Results on synthetic data suggest that our method exceeds the performance of established tools from bioinformatics [11].

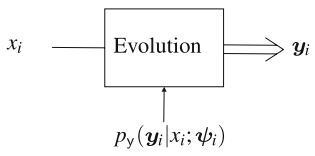

Fig. 8 Information transmission in evolution

5.5 Conserved non-genic sequences

Two to three years ago, when genomes from multiple sequences became available in high quality, the comparative methods revealed an unexpected feature of the DNA. It has been discovered that a lot of the conserved genome regions are non-genic, not coding for proteins [8,24]. These regions are believed to have important functions and are still poorly understood. If

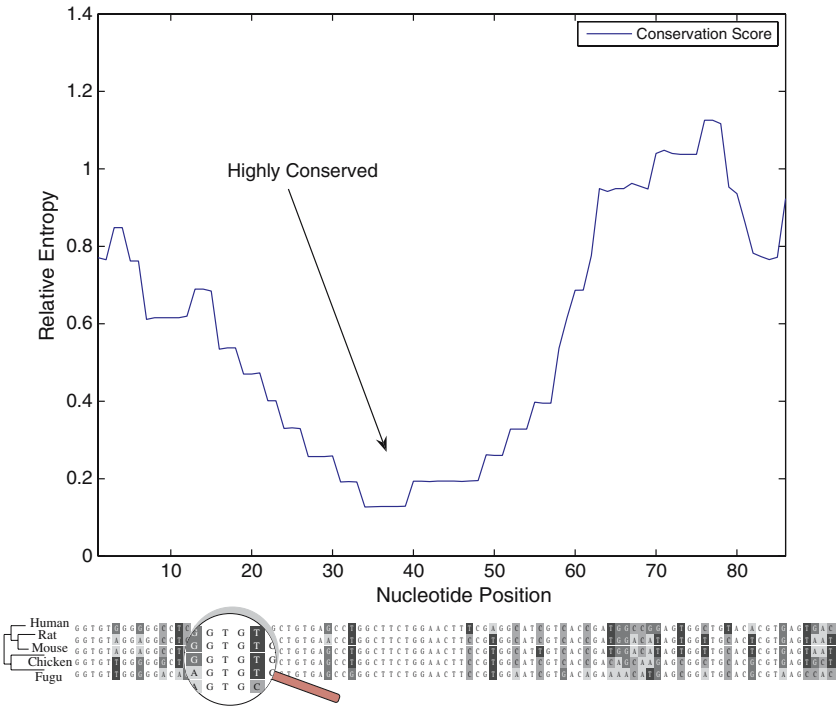

Fig. 9 *Top* the conservation scores indicating conserved regions. *Bottom* visualization of the respective genomic data, a small section of an alignment of the genomes of human, mouse, rat, chicken and fugu

an error correcting code exists on the genome sequence level, we expect the conserved non-genic regions to play a fundamental role in its implementation. Using our algorithm to identify conserved regions in the genome, future work will concentrate on the analysis of these conserved regions with respect to our hypothesis of an error correcting code on the genome level.

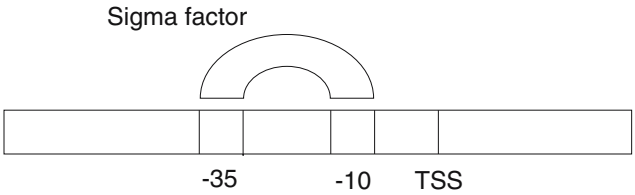

Fig. 10 Promoter detection by the sigma factor

6 Analogy between digital data transmission and transcription initiation

In digital data transmission the data is often divided into frames, whose header contains special patterns that indicate the beginning of the message in order to maintain synchronization. These patterns, the "sync words", need to be detected reliably by the receiver. Similarly, during transcription initiation—the first step of gene expression—the RNA polymerase has to recognize the promoter that indicates the beginning of a gene, see Sect. 2.4. In bacteria the RNA polymerase is directed to the promoter by the so called sigma factor. This sigma factor recognizes two short (six basepairs long) sequences separated by a spacer and positioned 35 and 10 basepairs (bp) before transcription start site (TSS). Therefore they are called the -35 and -10 regions. Hence, this process corresponds to a synchronization with two sync words in digital data transmission, see Fig. 10.

6.1 Choice of the sync words in binary and quaternary digital transmission

The sync words in digital data transmission have to be chosen such that they satisfy the following two conditions [2]: firstly, the probability of a random occurrence of the pattern in the data stream is to be minimized; secondly, the structure of the pattern should be such that the preceding symbols cannot yield a shifted sync word, as e.g. if the (binary) pattern is $+1+1+1+1+1+1$ there is a probability of 0.5 (assuming equally probable symbols) that it is followed by a $+1$ which may lead to a shifted synchronization. While the probability of a random occurrence does not depend on the sequence in case of independent symbols, the second condition is to be analyzed using the aperiodic autocorrelation function $\varphi_{ss}(\tau)$ of the sync word. $\varphi_{ss}(\tau)$ describes the similarity of a sequence $s = \{s_1, s_2, \ldots, s_l\}$ to itself for every shift $\tau \in [-(l-1); +(l-1)]$

$$\varphi_{ss}(\tau) = \sum_{m=1}^{l-|\tau|} s_m \cdot s^*_{m+|\tau|}, \tag{13}$$

where s^*_m denotes the complex conjugate of s_m. In order to minimize the probability of shifted synchronizations, the autocorrelation function of the sync word should have a narrow maxima at $\tau = 0$ and smallest possible values for $\tau \neq 0$ [16]. In general, the autocorrelation properties of a sequence are evaluated using the peak sidelobe (PSL)

$$PSL = \max_{\tau \setminus \{0\}} [\varphi_{ss}(\tau)], \tag{14}$$

which should be as small as possible to minimize the probability of false synchronizations.

6.2 Autocorrelation properties of *E.coli* promoter sequences

As mentioned before, transcription initiation corresponds to the process of synchronization used in digital data transmission, since two sync words—the promoter regions—need to be detected by the sigma factor. In order to gain more insights into promoter detection, we determine the autocorrelation properties of the −35 and the −10 promoter region in the bacterium *Escherichia coli* (*E.coli*) by adapting the autocorrelation function to the quaternary alphabet of nucleotides. Therefore, we have to redefine the product in (13) with respect to its biological meaning, i.e. such that it rates the effect of nucleotide matches and mismatches on the synchronization quality of the sequence. We rate an agreement of nucleotides by 1, a divergence of nucleotides by the negative value $-\frac{1}{3}$ (i.e. punishing mismatches with an overall weight of −1). This is done by introducing a mismatch score matrix D

$$D = \begin{array}{c} \begin{array}{cccc} \texttt{A} & \texttt{C} & \texttt{G} & \texttt{T} \end{array} \\ \begin{pmatrix} 1 & -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & 1 & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & 1 & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & 1 \end{pmatrix} \end{array} \begin{array}{c} \texttt{A} \\ \texttt{C} \\ \texttt{G} \\ \texttt{T} \end{array}, \tag{15}$$

and by replacing the product in (13) by the respective matrix values

$$\varphi_{ss}(\tau) = \sum_{m=1}^{l-|\tau|} D(s_m, s_{m+|\tau|}). \tag{16}$$

6.3 Results

The consensus (i.e. most frequently detected) sequences are TTGACA for the −35 region and TATAAT for the −10 region, respectively (see e.g. [13]). Figure 11 shows the autocorrelation functions of the two sequences. Calculation of the peak sidelobe for both promoter regions according to (14) results in

$$\text{PSL}_{-35} = \varphi_{ss}(|\tau| = 2) = 0,$$
$$\text{PSL}_{-10} = \varphi_{ss}(|\tau| = 3) = 1.67.$$

To rate the autocorrelation properties of the promoter sequences, we calculated the values of PSL for all $4^6 = 4,096$ possible nucleotide sequences of length 6. The mean value and the standard deviation of the resulting values are listed in Table 2.

It can be seen that the PSL of the −35 promoter sequence is highly below average, whereas that of the −10 promoter sequence lies above the mean value. In fact, only 1.15% of all possible sequences of length 6 have a better or equal PSL than the −35 region. Opposed to that, 79.37% of all sequences have a better or equal value of PSL compared to the −10 region. This fact suggests that nature employs a synchronization in two steps: firstly, the −35 region has to be detected out of all possible sequences with high accuracy to enable a reliable localization of the close-by transcription start site, see Fig. 12. In the second step, both regions are detected simultaneously, see Fig. 10, however, due to the synchronization conducted before, the sigma factor only needs to detect the −10 region out of around seven

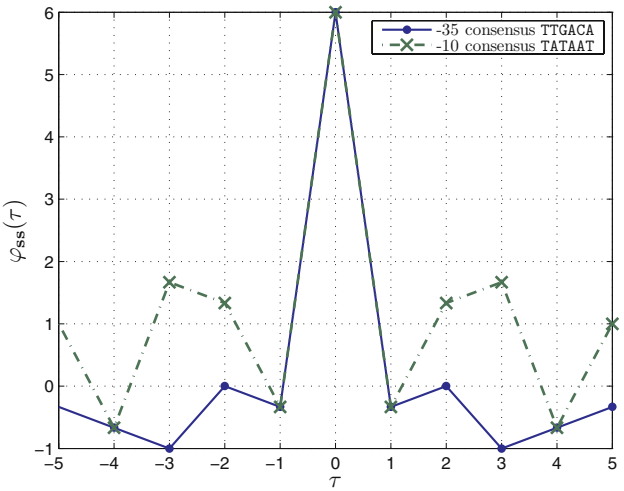

Fig. 11 Autocorrelation functions of −35 and −10 consensus promoter

Table 2 Mean and SD of PSL for all possible sequences of length 6

	PSL ratio
Mean	1.30
SD	0.76

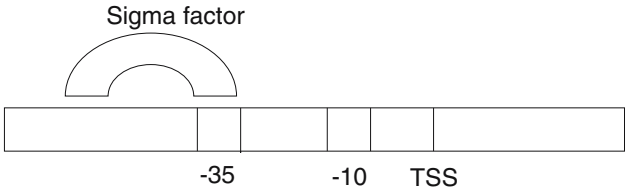

Fig. 12 Pre-synchronization during promoter detection

sequences based on the shape and limited deformability of the sigma factor that yield a variable spacing of 15 to 21 bp between the two promoter regions. Therefore, the sequence of the −10 promoter region is less important for synchronization. This brings up the conclusion that the two promoters might have evolved in a way to serve different tasks: while the −35 region is indispensable for indicating the close-by transcription start site and, thus, needs to have excellent synchronization properties, the sequence and structure of the −10 region seems to play a more important role during later steps of transcription initiation like DNA unwinding and opening, which require AT-richness (i.e. a high content of the nucleotides A and T) [23].

7 Conclusions

The newly available sequence data makes application of information theory to molecular genetics statistically feasible. Concepts like mutual information based distance measures combined with source coding can be applied to phylogenetic classification. Mutual information can be used for gene mapping of complex diseases. Additionally, communication theoretic models of information transmission can be used to search for error correcting codes in the genome or to gain better understanding of the molecular processes in the cell like the transcription initiation.

References

1. The international hapmap project (2003). Nature **426**(6968), 789–796. 1476–4687 (electronic) Journal Article Multicenter Study
2. Barker R (1953) Group synchronization of binary digital systems. In: Jackson W (ed) Communication theory, Butterworths, London pp 273–287
3. Battail G (2006) Introduction to Biosemiotics: information theory and error-correcting codes in genetics and biological evolution. Springer, Heidelberg
4. Cilibrasi R, Vitani PMB (2005) Clustering by compression. IEEE Trans Inf Theory 51(4):1523-1545
5. Cox T, Cox M (1994) Multidimensional scaling. Chapman & Hall, London
6. Dawy Z, Goebel B, Hagenauer J, Andreoli C, Meitinger T, Mueller JC (2006) Gene mapping and marker clustering using shannon's mutual information. IEEE/ACM Trans Comput Biol Bioinform 3(1):47–56
7. Dawy Z, Hagenauer J, Hanus P, Mueller JC (2005) Mutual information based distance measures for classification and content recognition with applications to genetics. In: Proceedings of the ICC 2005
8. Dermitzakis ET, Reymond A, Antonarakis SE (2005) Conserved non-genic sequences—an unexpected feature of mammalian genomes. Nature Rev Gene 6:151–157 URL http://www.dx.doi.org/10.1038/nrg1527
9. Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C, Antonarakis SE (2003) Evolutionary discrimination of mammalian conserved non-genic sequences (cngs). Science 302:1033–1035
10. Freeland S, Wu T, Keulmann N (2003) The case for an error minimizing standard genetic code. Orig Life Evol Biosph 33(4–5):457–77
11. Hanus P, Dingel J, Hagenauer J, Mueller J (2005) An alternative method for detecting conserved regions in multiple species. German conference on bioinformatics, Hamburg, p 64
12. Hayes B (1998) The Invention of the genetic code. Am Sci 86(1):8–14
13. Lewin B (2004) GENES VIII. Pearson Prentice Hall Upper Saddle River, NJ
14. Li M, Badger JH, Chen X, Kwong S, Kearney P, Zhang H (2001) An information-based sequence distance and its application to whole mitochondrial genome phylogeny. Bioinformatics 17(2):149–154. doi: 10.1093/bioinformatics/17.2.149
15. Lolle SJ, Victor JL, Young JM, Pruitt RE (2005) Genome-wide non-mendelian inheritance of extra-genomic information in arabidopsis. Nature 434(7032):505–509
16. Lueke H (1992) Korrelationssignale. Springer, Berlin
17. Margulies EH, Blanchette M, Program NCS, Haussler D, Green ED (2003) Identification and characterization of multi-species conserved sequences. Genome Res. 13(12):2507–2518 URL http://www.genome.org/cgi/content/abstract/13/12/2507
18. Mueller J, Bresch E, Dawy Z, Bettecken T, Meitinger T, Hagenauer J (2003) Shannon's mutual information applied to population-based gene mapping. Am J Hum Genet 73 (5 suppl) 610
19. Nirenberg MW, Matthaei JH (1961) The dependence of cell-free protein synthesis in e. coli upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci USA 47: 1588–602. 0027–8424 (print)
20. Sarkis M, Goebel B, Dawy Z, Hagenauer J, Hanus P, Mueller JC (2007) Gene mapping of complex diseases—a comparison of methods from statistics informnation theory, and signal processing. IEEE Sign Proc Magaz 24(1):83–90
21. Shannon CE (1940) An algebra for theoretical genetics. Ph.D. thesis, Massachusetts Institute of Technology, Dept. of Mathematics
22. Shannon CE (1948) A mathematical theory of communication. Bell Syst Techn J 27:379–423
23. Shomer B, Yagil G (1999) Long W tracts are over-represented in the *Escherichia coli* and *Haemophilus influenza* genomes. Nucleic Acids Res 27(22):4491–4500
24. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15(8):1034–1050 URL http://www.genome.org/cgi/content/abstract/15/8/1034

25. Ueda H, Howson J, Esposito L, Heward J, Snook H, Chamberlain G, Rainbow D, Hunter K, Smith A, Genova GD, Herr M, Dahlmand I, Payne F, Smyth D, Lowe C, Twells R, Howlett S, Healy B, Nutland S, Rance H, Everett V, Smink L, Lam A, Cordell H, Walker N, Bordin C, Hulme J, Motzo C, Cucca F, Hess J, Metzker M, Rogers J, Gregory S, Allahabadia A, Nithiyananthan R, Tuomilehto-Wolf E, Tuomilehto J, Bingley P, Gillespie K, Undlien D, Ronningen K, Guja C, Ionescu-Tirgoviste C, Savage D, Maxwell A, Carson D, Patterson C, Franklyn J, Clayton D, Peterson L, Wicker L, Todd J, Gough S (2003) Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. Nature 423(6939):506–511
26. Ureta-Vidal A, Ettwiller L, Birney E (2003) Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nat Rev Genet 4(4):251–262. 1471-0056 (print)
27. Wallace DC, Lott MT, Kogelnik AM, Brown MD, Navathe SB (1999) MITOMAP: a human mitochondrial genome database. URL http://www.dhgp.de/
28. Whelan S, Li P, Goldman N (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Gene 17(5):262–272