2

Bioinformatical Analysis of Point Mutations in Human Genome

Branko Borštnik and Danilo Pumpernik
National Institute of Chemistry Ljubljana, Slovenia

1. Introduction

1.1 The bioinformatical resources

The carrier of biological information is a DNA molecule that is packed in the cell nucleus. In the case of human species the genome consists of roughly three billion of base pairs that are packed into the chromosomes 1 through 22, and two sex chromosomes *x* and *y*. The genomic data acquisition witnessed a strong push in recent years. Ever more versatile sequencing technologies enable the sequencing laboratories to maintain a high throughput regime of work. Genomic data are freely available at several locations such as the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov) and European Molecular Biology Laboratory

(http://www.ebi.ac.uk/embl/) web sites. Also the aligned genomic sequences are available (Kuhn et al. 2009). The first draft of human genome was published in 2001 (Lander et al. 2001). A few years later Mikkelsen et al. (2005) produced the chimpanzee genome. Soon followed the rhesus genome (Gibbs et al. 2007) and also a great part of genomic sequences of orangutan and gorilla are available today.

The processes on the molecular level that are responsible for passing the genetic information between subsequent generations are prone to errors, causing the temporal modification of the genetic 'text'. In the primate branch of the phylogeny the error rate is approximately one point mutation per generation (Kumar and Subramanian 2002). Some changes eventually spread among a large part of the population, and the others die out. This happens within the period of approximately one million years (Myr) that is called the coalescence time. The differences between the human individuals that exist due to the mutations that did not yet reach their final destiny (i.e. extinction or spread over the entire population) represent the genetic polymorphism of a species. The richest amount of polymorphic data is available in the case of human species. We shall be only concerned with the so called single nucleotide polymorphism (*snp*). The word "single" in the *snp* phrase is to some extent misleading because the *snp* databases also contain entries with variations in the genetic text in the form of several consecutive variations of the nucleotide sequence of an individual with respect to the master sequence. The *snp* databases encompass more than 10 million entries (http://www.ncbi.nlm.nih.gov/snp). The problem is in validation of the *snp* data. Various

levels of validation are possible, but only a limited number of *snps* fulfill the most stringent validation criteria.

The split times of the human species with chimpanzee, orangutan, gorilla and rhesus macaque are in the range of 6 to 25 Myrs. Multiple sequence alignments reveal an order of magnitude of tens of millions of point mutations. This represents a decent basis for statistical evaluation of the evolutionary models. The human - chimpanzee sequence comparison (Mikkelsen et al. 2005) provides the most important set of data because their common ancestor lived 5 to 7 Myrs ago. Each nucleotide difference that is detected when comparing these two species is in the greatest extent the result of a single event. However, the double alignment does not uncover the directionality of the mutations (Jiang and Zhao 2006). In order to assess more complete information about the mutational processes one needs to align at least three genomic sequences. This means that one needs to go deeper in the evolutionary history by adding to the human - chimpanzee pair additional primate species, as mentioned above. Going deeper in the evolutionary history cannot be at no cost. The differences that are inferred from the multiple alignments do not need to represent the genuine replacements but a result of a combination of the events that might have taken place at the same DNA site. This is likely to occur at the hypermutable sites such as CpG dinucleotides (Fryxell and Moon 2005, Fryxell and Zuckerkandl 2000, Jabbari and Bernardi 2004, Ollila et al. 1996).

1.2 Nucleotide replacements

The single nucleotide replacement events can be categorized in several ways. There are 12 possible replacements - two pairs of transitions (A<=>G, C<=>T) and four pairs of transversions (A<=>C, A<=>T, G<=>C, G<=>T). The replacements are context dependent. On a coarse grained scale the contextual categories can be identified in terms of the DNA functionalities: transcribed and non-transcribed regions, regulatory regions and the remaining sequences. The regions are subjected to variable functional constraints and the mutational spectra would differ for various regions. One can also define the context on micro scale in terms of the flanking nucleotides (Siepel and Haussler 2004; Fryxell and Moon 2005; Zhao and Zhang 2006). One can choose only one nucleotide as the context defining element: the left or the right neighbor of the mutated site. In such a case the mutated site plus the nucleotide defining the context is a dinucleotide entity (Gentles and Karlin 2001). There are 16 distinct dinucleotides and the replacement counts and the replacement probabilities can be represented by 16x16 matrices. Two kinds of matrices will be the subject of our interest. The **A** matrices will represent the replacement counts and the **W** matrices will represent the replacement probabilities. The other possibility is that the context is defined in terms of right and left neighbor. In this case the above mentioned matrices are of rank 64. However, one does not need to deal with all 64x64=4096 matrix elements. If the 12 above mentioned transitions and transversions are inserted within 16 possible left/right nucleotides defining the context one obtains altogether 192 replacement types that need to be examined in terms of their frequencies of appearance and their probabilities.

The counts can be extracted from the genomic alignments and single nucleotide polymorphism data. In order to unravel the evidence regarding the compositional changes one should gain access to the directionality of the replacements, meaning that when two

different nucleotides are found in two genomes at a certain site one should be able to tell, which is the ancestral and which is the newly replaced nucleotide. The identity of the ancestral sequence elements can be determined by the analyses of multiple sequence alignments. The majority principle is usually applied. This means that the nucleotide that is found most frequently at a certain site is supposed to be the ancestral one. In the case of multiple events taking place at a single site the majority principle does not lead to proper results, which can be corrected by computer simulation.

We shall adhere to the convention that the A_{ij} element represents the number of j to i replacements. On the basis of the **A** matrix one can define the replacement probability matrix **W** following the simple philosophy that the number of replacement events is equal to the number of the sites multiplied by the replacement probability

$$A_{ij} = W_{ij} N f_j . \quad (1)$$

$N=\Sigma_i\Sigma_j A_{ij}$ denotes the total number of dinucleotides and f_j are the elements of the dinucleotide composition vector ($f_j = \Sigma_i A_{ij} / N$). The **W** matrix has the property that each column sums to unity:

$\Sigma_i W_{ij} = 1$.

1.3 Mutations produced by the replication slippage

A rather potent mechanism of DNA modification is the modification of the length of short tandem repeats or microsatellites (Cox and Mirkin 1997, Borštnik and Pumpernik 2002, 2005, Borštnik et al. 2008). The human genome contains approximately 2% of sequences having the form of short tandem repeats of nucleotides, dinucleotides, trinucleotides and so forth. In the DNA replication process the repetitive parts of DNA sequences are likely to be incorrectly reproduced. The repeats can become elongated or shortened. This is because the two complementary DNA strands retain the complementarity also in the case when one or the other strand slips forward or backward for an integer number of repeat units. In a series of subsequent cell divisions a tandem repeat can thus lead to a substantial elongation or, in the other extreme, may even lead to the disappearance of the repeat. The high mutability of repeats makes them usable in genotyping purposes. Microsatellite markers exhibit mutation rates that exceed the average mutation rates by two orders of magnitude. In this work we shall put under scrutiny whether the nucleotide replacement process and the replication slippage mechanism produce comparable densities of mutational changes.

1.4 The distribution of mutations along the chromosomes

Since stochasticity is an essential component of the mutational processes one can expect that the locations of the mutated sites will be randomly distributed along the DNA sequence. According to our opinion it is worth to study in detail the spatial distribution of the mutated sites. We shall be concerned with three types of point mutations: nucleotide replacements produced by the genomic and *snp* mutations and replication slippage events. There are several possible realizations of mutation density studies. One can look for the differences between genomic and *snp* nucleotide replacements and replication slippage mutations. Further, one can compare the mutation densities occurring in the human genome to those

occurring within the genomes of other primate species. This can be done if one can trace the directionality of the mutations. By aligning several primate species one can first determine the ancestral state of the sequence and then one can identify the species that was subjected to the mutation.

1.5 The neutrality hypothesis

It is not easy to explain the Darwinian principle of the survival of the fittest in terms of the changes taking place at the DNA level. The idea that the mutations can be divided into two classes – the ones giving a positive contribution to fitness, and the others reducing the fitness of the organisms is an oversimplification. Kimura (1968) has shown that the majority of changes at the molecular level are neutral and do not affect fitness. Stochasticity is the main characteristic of molecular evolution. The changes are generated at random times and random positions along the chromosomes. The driving force of the mutations is the non-ideality of the DNA replication apparatus. If the selective coefficient belonging to the change that the mutation produces is small enough, the mutated variants would eventually spread among the population since there is no impact upon the functioning of the organisms. Studies of the patterns of neutral mutations can offer an insight into the cellular machinery that is responsible for DNA replication.

1.6 Scope of the work

In what follows we present new results on the mutational propensities in the human lineage. In the next section the results obtained by a simple four parameter model of nucleotide replacements will show that the mutational changes are amenable to statistical predictions. In the next section the model is expanded to a multi-parameter form and it is shown that the agreement between the model and the results of the alignment of natural sequences become very close. The mutational analyses are then expanded in two respects. The comparison between the processes leading to the diversity and divergence is performed on the ground of the nucleotide replacements. Further, two competing mutational mechanisms are compared: the nucleotide replacements and the replication slippage mechanism. The question is also addressed how to search the genetic basis of human phenotypical characteristics.

2. Results

2.1 Nucleotide replacements

Let us first present the results obtained by a simple four parameter model of nucleotide replacements. The model is defined as follows. A nucleotide sequence with the 60:40 A+T:C+G composition is subjected to a process of one-nucleotide replacement at random positions in such a way that transitions are four times more probable than transversions. The CpG dinucleotides are taken to be five times less frequent and two times more mutable than the average dinucleotides. A computer simulation procedure was set up. An artificial sequence was randomly subjected to the nucleotide replacements. After enough replacements have been accumulated to reproduce the human/chimpanzee split the counts were performed on the level of trinucleotides and the results were compared with the results of the human - chimpanzee genomic alignments. The results are presented in Fig. 1,

where the logarithm of human/chimpanzee replacement counts are plotted along the horizontal axis and the model values along the vertical axis. The points are scattered along the $y=x$ line. The four groups of points belong to $\Delta n=0,1,2,3$ substitutions per trinucleotide. Each group is dispersed and divided into several subgroups. The $\Delta n=0$ group located at the upper right corner of the figure represents the degree of reproduction of the genome composition in terms of trinucleotides. It shows that a two parameter (the C+G:A+T ratio and the degree of CpG depletion) model reproduces reasonably the DNA composition in terms of trinucleotides. The points in the form of full circles are divided into two subgroups on the basis of CpG content. The eight trinucleotides in the form XCG and CGX are separated from the remaining trinucleotides due to their strong depletion. The other three groups ($\Delta n=1,2,3$) are further partitioned on the basis of transition/transversion differences and CpG supermutability. The positions of $\Delta n=3$ points in the form of empty triangles are systematically deflected below the $y=x$ line which means that the multiple replacements at the neighboring sites are not as independently occurring as a model of independent mononucleotide events would predict.

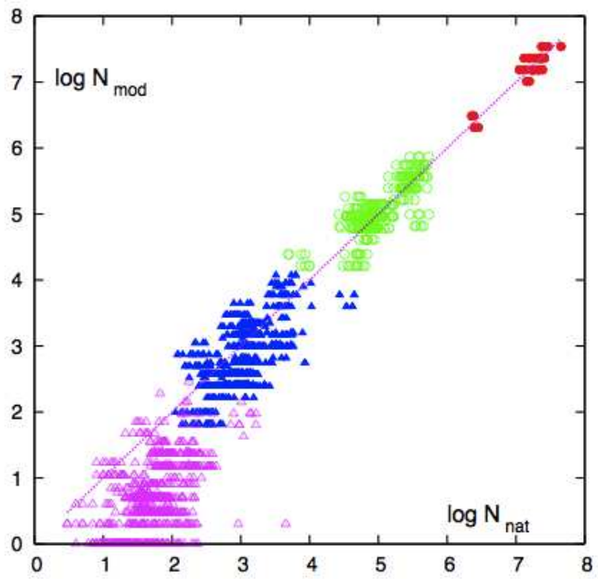


Fig. 1. The comparison between the human - chimpanzee replacement counts based on a simple four parameter model (vertical direction - N_{mod}) and the results based upon the human - chimpanzee genomic alignment (horizontal direction - N_{nat}) in trinucleotides. The four symbols (full circle, empty circle, full triangle, empty triangle) refer to Dn=0, 1, 2 ,3 replacements within a trinucleotide frame.

2.2 Multi-parameter dinucleotide replacement model

In general one can construct a multi-parameter nucleotide replacement model. Hwang and Green (2004) presented the most elaborate model by dividing the genomic sequences in several classes according to their functional role. The context was taken into account with left and right neighbor and the replacement probabilities for all possible pairs of trinucleotides were optimized in order to match the results of multiple genomic alignments. Only those trinucleotide pairs were taken into account where the left and right nucleotides

are identical while the middle one is being replaced. Also several other authors (Arndt and Hwa (2005), Baele et al. (2008), Duret and Arndt (2008), Lunter and Hein (2004), Borštnik and Pumpernik (to be published)) were trying to infer the replacement probabilities. It turns out that the results are pretty elusive. There are several circumstances that render the parameterization of the mutational process difficult. It is hard to attain the statistical significance. A particular class of mutational changes can easily be defined in such a way that, when it is sampled within the appropriate database becomes too narrow to guarantee a satisfactory statistics. In this work we would like to present an approach with the classification of the mutational changes that does not follow the standard practice in order to gain a profit in the form of the reduction of statistical fluctuations. We are halving the number of the nucleotides defining the context and perform the analyses in the space of dinucleotides. In this case the number of unknown replacement probability parameters is reduced to the number of the entities defining the context (4 in our case) multiplied with the number of distinct type of replacements (four transitions and 8 transversions, what makes 12 replacements), thus 48 unknowns. In order to further improve the statistics we have chosen the largest set of nucleotide sequences possible: the complete human genome aligned with the counterparts of chimpanzee, orangutan and rhesus genomic sequences. The number of nucleotides aligned was close to one billion.

The simulation procedure was performed in order to determine the elements of the replacement probability matrix and to retrieve the replacement count matrix that is free of superposition errors. This goal was achieved by simulating the replacement process on an artificial starting DNA sequence. The simulation was run along the dendrogram in the form of four branches. The branch representing the outgroup species (rhesus or orangutan) extends from the common root to the present. Its length is denoted by T_r or T_o , respectively. The human/chimpanzee common ancestor branch extends from the common root to the point of human/chimpanzee split and its length is denoted by T_{hp}. The replacements were generated using the random number generator uniformly on all the branches according to the probabilities defined by the mutation probability matrix **W**. The **W** matrix was expressed as

$$W_{ij}=a_{ij} A_{ij}/(Nk_j f_j) \quad (2)$$

where A_{ij} and f_j elements were taken from the results of the triple alignments and a_{ij} and k_j are the correction coefficients subjected to the optimization procedure. In a hypothetical ideal case when the **A** matrix would represent genuine replacement numbers free of superposition errors, all the a_{ij} and k_j coefficients would be equal to unity. In our case the a_{ij} and k_j coefficients were determined by the following procedure: After each simulation round the three artificial sequences were aligned and the resulting $\mathbf{A}^{(t)}$ matrix was compared with the corresponding $\mathbf{A}^{(t)}$ matrix obtained when aligning the natural sequences. The difference between the two matrices was minimized by optimizing the correction coefficients by a brute force Monte Carlo procedure. Random variations were performed and the variations not improving the matching of the two matrices were discarded. The variational procedure in such an enormous parameter space as it is spanned by the entire set of the correction coefficients is not easy to carry out. It is expected that the main source of the superposition events are the replacements caused by CG dinucleotide decay. Therefore the Monte Carlo procedure was conducted in such a way that the parameter subspace associated with the CG decay was given more attention than the remaining regions of the

parameter space. The end result of the simulation procedure were the **W** matrix and the matrix $\mathbf{A}^{(d)}$ of direct counts of the dinucleotide replacements that was free of deformation due to multiple replacements taking place at the same site.

In Fig. 2 the pairs of $\mathbf{A}^{(t)}$ matrix elements are plotted in such a way that the i <= j replacement counts are used as the x coordinate and the j <= i value as the y coordinate. The two points belonging to the same i,j pair are positioned symmetrically with respect to the y=x line and the distance between them represents the measure of the asymmetry. The dinucleotides connected by the strand symmetry are positioned approximately at the same place. The data plotted in Fig. 2 were extracted from the human/chimpanzee/rhesus (*hpr*) alignment. We can see that the majority of the replacements exhibit an appreciable degree of asymmetry. When these data are compared to the human/chimpanee/orangutan (*hpo)* case it turns out that the asymmetry is independent of the choice of the triple alignment sources. Nearly all the pairs of points belonging to *hpr* and *hpo* coincide to a rather high extent. Only the points corresponding to the CG <=> CA/TG replacements are in disagreement. The

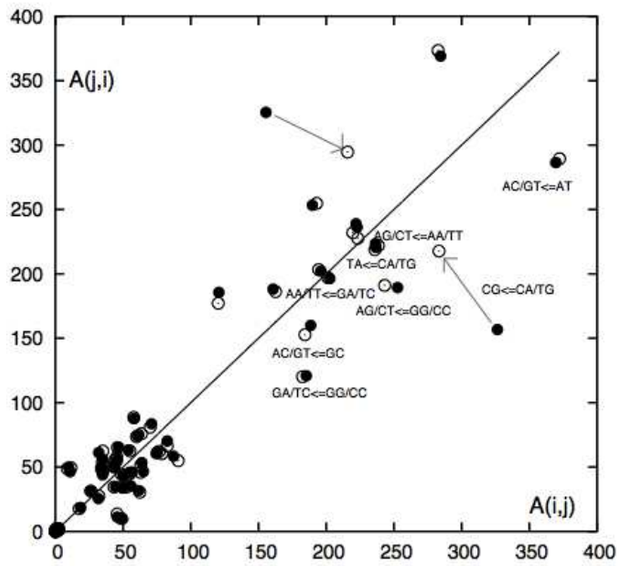


Fig. 2. The dinucleotide replacement counts presented in such a way that each pair of counts running in opposite directions is represented by a point in A_{ij}, A_{ji} plane. If the two counts are equal the corresponding replacement would be located along the y=x line. The distance from y=x line is a measure of the directionality asymmetry. The two replacement pairs connected by the strand symmetry is mapped as a pair of points mirroring across the y=x line. Bullets correspond to the replacement counts obtained from human/chimpanzee/rhesus triple alignments ($\mathbf{A}^{(t)}$) and circles to the direct replacement count ($\mathbf{A}^{(d)}$) generated in the simulation procedure. In nearly all the cases the bullets and circles coincide. A significant discrepancy takes place in the case of CG <=> CA/TG replacements where a pair of arrows indicate the extent of the superposition error inflicted upon the counts retrieved by the triple alignments. Only the replacements of the transition type below the y=x line are interpreted. The interpretation is also valid for the points that are mirrored across the y=x axis, provided that the arrows symbolizing the replacement directions are reversed. The units are arbitrary.

asymmetries of the CG <=> CA/TG replacements are larger in the case of *hpr* than in the case of *hpo*. Such a disagreement leads us to the idea that the CG <=> CA/TG replacement asymmetries are the artifact of the procedure by which they were detected. The process to be blamed is the superposition of multiple replacement events at a single site which was resolved by the simulation of the replacement process.

The optimal agreement between the natural and simulated $\mathbf{A}^{(t)}$ replacement count matrices was achieved with the branch length ratio $T_{hp}/T_r = 0.8$ and $T_{hp}/T_o = 0.53$. The majority of the resulting correction factors emerged from the optimization procedure within the interval 0.96 to 1.04. The following values emerged outside this interval $a_{ij}=1.3$; $a_{ji}=0.9$ and $k_j=1.35$ for j=CG; i=CA/TG. The final correlation coefficient between the two (natural and model) $\mathbf{A}^{(t)}$ matrices was equal to 0.9995 when simulating the *hpr* case. The two direct count matrices $\mathbf{A}^{(d)}$ belonging to the *hpr* and *hpo* case exhibit small differences. The extent of corrections that emerge from the simulation procedure is seen in Fig. 2. The two arrows mark the position of CG<=>CA/TG replacements in the A_{ij}/A_{ji} plane corresponding to the *hpr* case.

The dinucleotide replacement probabilities W_{ij} that resulted from the optimization procedure are presented in Fig. 3 in a similar way as the A_{ij} counts are presented in Fig. 2.

The location of the two points belonging to CG decay are located outside the margins of the figure and their position is indicated by the arrows. The probability of CG decay is for one order of magnitude above the average probability of the transition type replacements. The asymmetries of the dinucleotide replacements do not need to exhibit the same direction in the W_{ij} and A_{ij} cases because the factor f_i/f_j in the relation $W_{ij}/W_{ji} = (A_{ij}/A_{ji})(f_i/f_j)$ can reverse the directionality of the two pairs of i,j values.

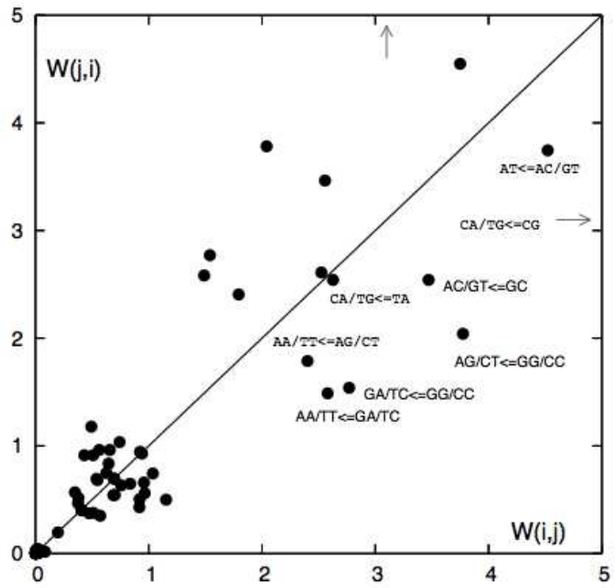


Fig. 3. The dinucleotide replacement probabilities W(i,j) presented in the same way as the A(i,j) counts in Fig. 2. Note that the italicized dinucleotide pairs exhibit the opposite directionalities in comparison with the A(i,j) values. The CA/TG<= CG replacement probabilities are located in the direction indicated by the two arrows at x=21 (horizontal arrow) and and y=21 (vertical arrow). The units are arbitrary.

The deviation from reciprocally equilibrated replacements between the dinucleotide pairs are statistically the most significant in the cases where the two dinucleotides differ at one place - and the difference is of the transition type. The dinucleotide pairs of this type can be grouped into three clusters (Figs. 4 and 5). The first cluster (Fig. 4) comprises the dinucleotides GG/CC that can be replaced by the dinucleotides GA/TC or AG/CT, which can be further replaced by AA/TT dinucleotides. The next two clusters (Fig. 5) comprise the four palindromic nucleotides of which AT and GC can be replaced by AC/GT, while TA and CG can be replaced by CA/TG dinucleotides. Also the asymmetries of the three clusters of the replacements are presented in Figs. 4 and 5. It is evident that the transition components of the fluxes are not equilibrated. In order to see whether the replacement process is running in a steady state or not, one should examine the **A** and **W** matrices in their entirety. Both $\mathbf{A}^{(d)}$ matrices, in the *hpr* and *hpo* case clearly show that in the case of AC/GT, TA and CA/TG dinucleotides are increasing their share in the replacement process, while CC/GG, AG/CT and AT are losing their share in dinucleotide population.

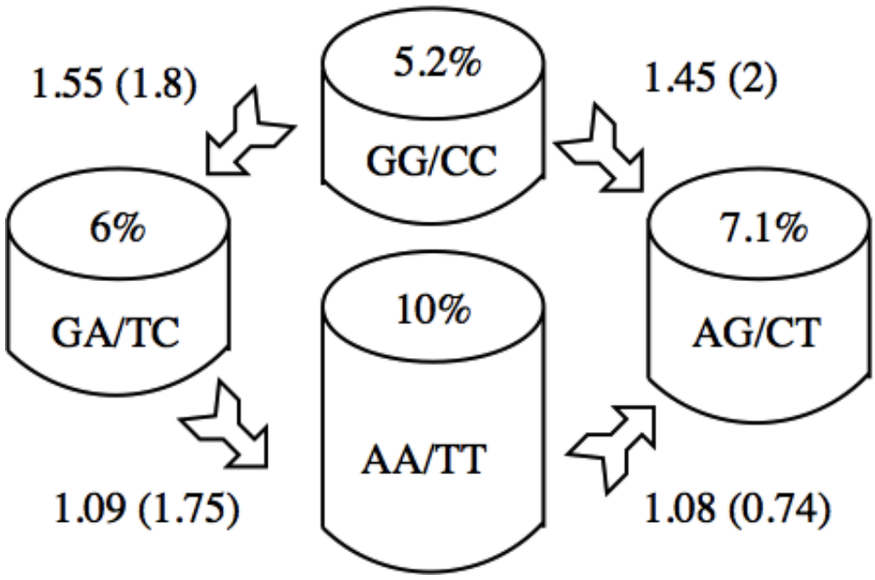


Fig. 4. The asymmetries of the GG/CC, GA/TC, AG/CT replacements. The height of the barrels is proportional to their composition (given in percents at the top). For each replacement the arrow shows the predominant directionality. The numbers beside the arrows give the asymmetry factor - the $A_{ij}^{(d)}/A_{ji}^{(d)}$ ratio. The numbers in parentheses give the corresponding Wij/Wji ratios.

2.3 The density of mutations: Diversity versus divergence comparison

The genetic diversification within the human population appears on the account of two mechanisms – both having essential stochastic component – point mutations and genetic recombination in the process of meiosis. The former process generates new varieties and the latter one generates new combinatorial realizations of the genetic differences. Nucleotide sequencing reveals the differences between the individuals and make them available for bioinformatical processing. Roughly 10^7 single nucleotide polymorphisms are known today what means one to ten polymorphic sites per thousand nucleotides.

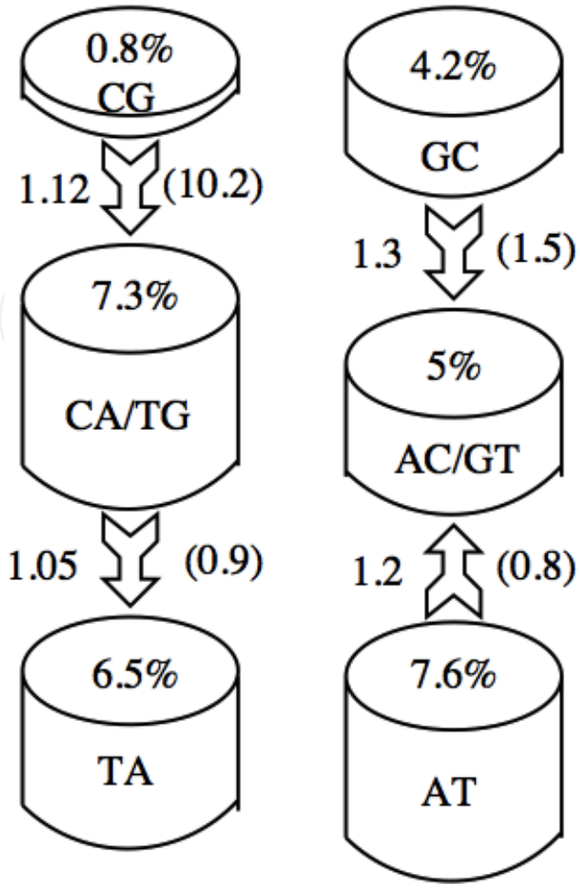


Fig. 5. The asymmetries of the GC, AC/GT, AT, CG, CA/TG and TA replacements. See also the caption of Fig. 4.

Also the differences between the master sequences of closely related diverging primate species exhibit a comparable densities of point mutations as the above mentioned intraspecies diversities. One can compare the two types of mutabilities in several ways. The standard approach is to compare the nucleotide replacement matrices. Our preliminary results (Borštnik and Pumpernik - in preparation) show that there exists a certain difference between the replacement matrices. In this work we focus our attention on the question as to how the two types of mutations are distributed along the chromosomes. In Figure 6 the distribution of genomic and *snp* mutations is presented for a typical segment located on the first chromosome. The density of mutations is proportional to the slope of the line on which lies the sequence of points representing the counts of nucleotide replacements. We can see that, roughly speaking, there exist chromosomal segments with diverse densities of mutations. This is valid for *snp* and for the genomic mutations. However, the densities of the two types of mutations do not necessarily fluctuate in phase. A more detailed information regarding the distribution of the replacement sites is presented in Figure 7. The two horizontal directions in the plot refer to the number of two types of the replacements per 1000 nucleotides. The direction marked with "gen" refers to the density of genomic mutations and the other horizontal coordinate corresponds to the density of the *snp* replacements. The values along the vertical direction represent the number of 1000 bp

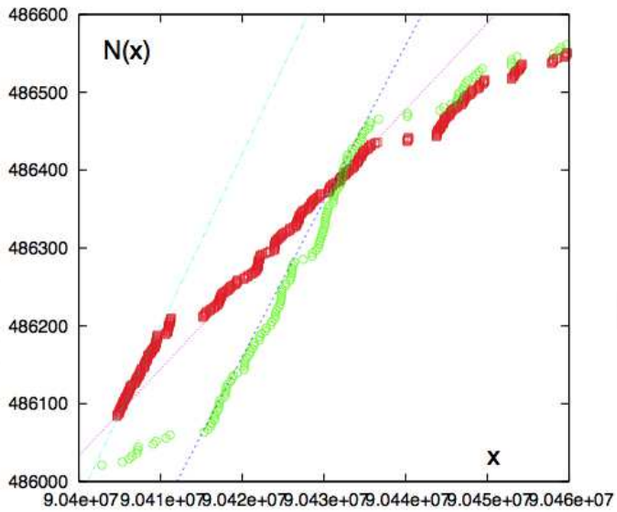


Fig. 6. The comparison of counts of genomic and *snp* type of nucleotide replacements. Horizontal variable x represents the coordinate (running nucleotide number (from 1 to 250 million)) along the first chromosome, N(x) is the cumulative number of human – chimpanzee nucleotide replacements within the interval [0-x] (red square symbols). Green circles represent the *snp* counts as a function of the chromosome coordinate. The *snp* counts are multiplied by a factor 3.827 in order that only one series of values suffice to be displayed along the vertical axis. The nucleotide replacement densities are proportional to the slope of N(x) plot. One can identify segments with well defined densities, that vary from segment to segment. The straight lines are plotted to guide the eye.

segments possessing the characteristics defined by the two horizontal coordinates. The distribution is close to what one would expect. The plot in two dimensions exhibits its highest values close to zero density and not at the average value that is located at higher values of the densities. The essential message of Fig. 7 is the following. In spite of the fact that the genomic and *snp* replacements exhibit strong fluctuations in mutation density and that the fluctuations are apparently out of phase, the two dimensional distribution of the two densities exhibits normal unskewed properties.

Besides the comparison of genomic and *snp* replacements we also performed the comparison between the nucleotide replacement type of point mutations and replication slippage type mutations. Short tandem repeats (*str*) with the monomer lengths 1 and 2 were first located within the human sequence. In the next step the human – chimpanzee *str* counterparts with unequal lengths were detected in the human/chimpanzee alignment. The result was a list of chromosomal coordinates of replication slippage type of mutations. This list can be compared with the list of genomic nucleotide replacement type mutations. The question can be posed again whether there is a noticeable correlation in the distribution of the mutations of various classes. In Figure 8 the mutation counts are plotted as a function of the chromosome coordinate. The densities of three kinds of mutations are presented: mutated *strs* with monomer lengths 1 and 2 and single nucleotide replacement mutations that were already presented in Fig. 6. We can see again that the densities of the three kinds of mutation do not oscillate in phase. The strongest fluctuations are present in the

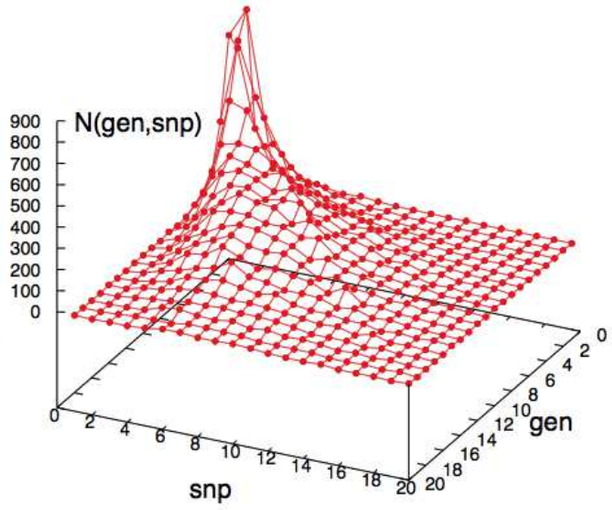


Fig. 7. A three dimensional plot of the genomic and *snp* type nucleotide replacement densities. The coordinates of the points in two horizontal directions have the following meaning: number of genomic replacements per 1000 nucleotides for the "gen" direction and number of *snp* nucleotide replacements per 3800 nucleotides for the "snp" direction. The heights of the points represent the number of segments having the densities specified by the two horizontal coordinates.

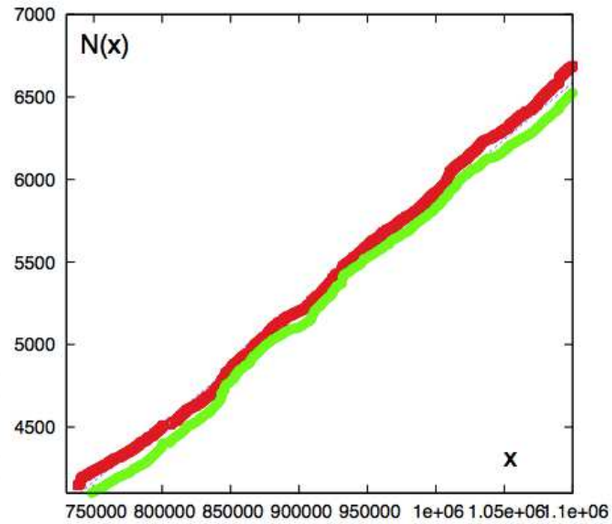


Fig. 8. Red squares, green circles and blue triangles (hidden behind the red squares) represent the counts of nucleotide replacements and repeat elongation/shortening of mononucleotide and dinucleotide repeats, respectively, as a function of the first chromosome coordinate x.

distribution of *strs* composed of dinucleotides. This is to be expected. There is a significant difference between the mononucleotide and dinucleotide repeats. Mononucleotide repeats of the type C_n or G_n are very rare. The majority of mononucleotide repeats are of the form A_n or T_n and more than one half of them have their origin in retroposed segments of

retroposons with a polyadenyl tail. After these elements are retroposed they begin with their mutational dynamics. The dinucleotide repeats, on the other hand, are supposed to be the result of a pure replication slippage dynamics and therefore the densities of their mutations exhibit a different pattern than the mononucleotide repeats.

2.4 The density of mutations: Human versus non-human mutations

As the last case, let us present the analysis of the differences between the densities of the mutations that occurred in human lineage on one hand and the densities of the mutations that occurred in other primate species. We analyzed quadruple alignment of human, chimpanzee, orangutan and gorilla genomic sequences. Suppose that a site is populated by x,y,z,w nucleotides at their respective human, chimpanzee, orangutan and gorilla sequence. In the majority of cases all four nucleotides are identical, indicating that the site was not subject to mutation. The cases where the diversity of x,y,z,w goes beyond two unequal nucleotides are rare and were ignored. What remains are the cases where a site is populated by two nucleotides, say x and y. There are 6 realizations of (2x,2y) case and 4 realizations of (3x,y) case. We considered only the cases y,x,x,x and x,y,x,x. In the first case it is plausible to conclude that the mutation took part in human lineage and in the second case chimpanzee lineage can be supposed to be hit by the nucleotide replacement. In our earlier study (Borštnik and Pumpernik - in preparation) we produced the trinucleotide 64x64 replacement matrices for the cases where only the middle nucleotide is allowed to be replaced. Using the above mentioned criterion to determine the directionality of the replacement we have shown that no significant difference exists between the nucleotide 64x64 replacement matrices taking place in human lineage and chimpanzee lineage. Also the results based upon the study of sequential distribution of the nucleotide replacements shown in Fig. 9 corroborate the notion that on average there is no appreciable difference between the mutation densities in the two species. The most important information emerging from Fig. 9 can be extracted by comparing the mutation densities in human and chimpanzee species. To

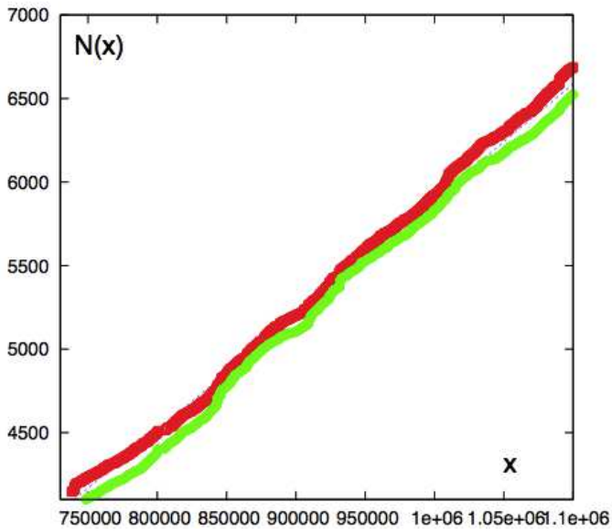


Fig. 9. Red squares and green circles represent the counts of nucleotide replacements taking place in human and chimpanzee species, respectively, as a function of the first chromosome coordinate x.

some extent the two densities go hand in hand, but the deviations of this rule are appreciable. This means that after the human – chimpanzee split the mutational process in each genome ran with comparable pace in orthologous regions. Obviously in the course of time species-specific mutation patterns came into the existence. Detecting and scrutinizing such regions can provide vital information about the human ascent.

3. Discussion

The purpose of this work is to contribute new findings in the field of the studies of point mutations in human lineage. This topics is of paramount importance because at the moment a greater part of questions concerning the correspondence between genotypic and phenotypic mutational changes remain unsolved. A conservative scenario as to how to figure out the meaning of the genomic differences should point towards the 22.000 primate genes and, of course towards their regulatory regions. The differences between the coding regions of the primate genes are known (Jordan et al. 2006, Goldstein and Pollock 2006), but the exact identity of the regulatory sequences is still missing (Tuch et al. 2008) and one does not know exactly what to search for in the genomic comparison between human and other primates. This means that the studies of the genomic differences between human and other great apes are welcome even if the studies are not directed to a specific gene product. One can follow the philosophy due to which one decides to accumulate and analyze mutational data on large scale. Doing this one should carefully evaluate the statistical errors. A simple guideline to estimate the uncertainties can be deduced from the theory of radioactive decay for which we know that if one measures N events the relative uncertainty of the quantities that are deduced from this measurement is 1/sqrt(N). This means that when counting, for instance, the number of replacements of a pair of dinucleotides or trinucleotides and finds N occurrences, the relative uncertainty of the corresponding replacement probability will be again 1/sqrt(N). In order to achieve 1% accuracy N should be of the order of magnitude of 10^4. When determining the dinucleotide mutation probability matrix the number of essential replacement classes is of the order of magnitude of 100. To accumulate 10^4 replacements per replacement class one needs to have at least 10^6 replacement events. Taking into account one percent difference between distinct primate sequences the body of aligned sequences should encompass a sequence length of at least 10^8 bp. Such an amount of aligned sequence material is rarely reported in the literature. Usually the analyses are performed on the samples containing a set of pseudogenes, introns or other sequence samples with low functional constraints (Zhang and Gerstein 2003). Our approach in the studies of dinucleotide replacement was based upon the premise that more than 95% of genetic sequences (Davydov et al. 2010) is subjected to low functional constraints and therefore we decided to align the entire set of genomic sequences.

To return to the question regarding the genetic basis of human phenotypic abilities let us discuss the results presented in Fig. 9. The general picture supports the notion that the two species are pretty equilibrated in the nucleotide replacement densities. The two replacement counts grow nearly at the same pace as a function of chromosome coordinate. However, it is possible to find the difference in details. There are regions where human replacements are more dense than the chimpanzee counterparts. Such regions are the candidates for coding the human specific traits. There is of course a long way to go to attain an in-depth view as to

how the human specific traits are coded. One could combine the results of studies of nucleotide replacement densities with the studies of positive or accelerated evolution (Wagner 2007) of protein coding regions. Under normal circumstances the mutabilities of the protein coding regions is an oscillatory function with period 3 – what is the codon length. The third position in a codon is usually, due to the codon degeneracy, subjected to lower constraints than the other two positions and exhibits higher mutability. The replacements taking place at synonymous sites are neutral and the replacements at the nonsynonymous sites produce amino acid changes. It is not easy to detect the deviations of the expected mutational patterns that can indicate the presence of an accelerated evolution. One can envisage the following scenario: The candidates for the accelerated evolution can be sought within the regions where the human genome exhibit mutation densities that exceed the mutation densities of other primate species. Within these regions one would then look for the genetic changes that shaped the properties of human species.

4. Conclusions

The purpose of this work is to contribute new findings in the field of the studies of point mutations in the human genome. We show that the most appropriate approach is to analyze the results of multiple alignments of primate genomic sequences. Several methods are presented how to carry out the analyses. The construction of dinucleotide replacement probabilities provides us with the information about the details of the nucleotide replacements. The role of CpG dinucleotide is unveiled and it is shown how the hypermutability of CpG dinucleotide influences the influx and outflux of all the other dinucleotides. Another aspect of the mutational processes, which is put under scrutiny in this work, is the study of sequential distribution of point mutations. The properties of three types of mutations is compared: nucleotide replacements involved in diversity and divergence processes and replication slippage events. It is also shown how the density of nucleotide replacement events along the chromosome can unveil the sites where human - specific traits can be coded.

5. Acknowledgment

This work was financed by the Slovenian Research Agency.

6. References

Arndt, P F & Hwa T (2004) Regional and time-resolved mutation patterns of the human genome. *Bioinformatics* 20:1482-1485

Borštnik, B & Pumpernik D (2002) Tandem repeats in protein coding regions of primate genes. *Genome Res.* 12:909-915

Borštnik, B & Pumpernik D (2005) Evidence on DNA slippage step-length distribution. *Phys. Rev.* E71:031913;1-7

Borštnik, B; Oblak, B & Pumpernik, D (2008) Replication slippage versus point mutation rates in short tandem repeats of the human genome. *Mol. Genet. Genomics* 279:53-61

Cox, R & Mirkin, S M (1997) Characteristic enrichment of DNA repeats in different genomes. *Proc. Natl. Acad. Sci.* USA 94:5237-5242

Davydov, E; Goode, D; Sirota, M; Cooper, G; Sidow, A; et al. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* 6:e1001025 (13pages)

Fryxell, K J & Moon W-J (2005) CpG mutation rates in the human genome are highly dependent on local GC content. *Mol. Biol. Evol.* 22:650-658

Fryxell, K J & Zuckerkandl, E (2000) Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol.* 17:1371-1383

Gentles, A J & Karlin, S (2001) Genome scale compositional comparisons in eukaryotes. *Genome Res.* 11:540-544

Gibbs, R A; Rogers, J; Katze, M G et al. (181 co-authors) (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316:222-234

Goldstein, R A & Pollock, D D (2006) Observations of amino acid gain and loss during protein evolution are explained by statistical bias. *Mol. Biol. Evol.* 23:1444-1449

Jabbari, K & Bernardi, G (2004) Cytosine methylation and CpG, TpG(CpA) and TpA frequencies. *Gene* 333:143-149

Jiang, C & Zhao, Z (2006) Directionality of point mutation and 5-methylcytosine deamination rates in the chimpanzee genome. *BMC Genomics* 7: 316 doi:10.1186/1471-2164-7-316

Jordan, I K; Kondrashov, F A; Adzhubel, I A; Wolf, Y I; Koonin, E V; Kondrashov, A S & Sunjaev S (2006) A universal trend of amino acid gain and loss in protein evolution. *Nature* 433:633-638

Kimura, M (1968) Evolutionary rate at the molecular level. *Nature* 217:624-626

Kuhn, R M; Karolchik, D; Zweig, A S; et al. (22 co-authors) (2009) The UCSC Genome Browser Database: update 2009. *Nucleic Acid Res.* 37:D755–D761

Kumar, S & Subramanian, S (2002) Mutation rates in mammalian genomes. *Proc. Natl. Acad. Sci. USA* 99:803-808

Lander, E S; Linton, L M; Birren, B; et al. (249 co-authors) (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860-921

Michel, C J (2007) Evolution probabilities and phylogenetic distance of dinucleotides. *J. Theor. Biol.* 249:271-277

Mikkelsen, T S; Hillier, L W; Eichler, E E; et al. (67 co-authors) (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69-87

Ollila, J; Lappalainen, I & Vihinen, M (1996) Sequence specificity in CpG mutation hotspots. *FEBS Lett.* 396:119-122

Siepel, A & Haussler D (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* 21:468-488

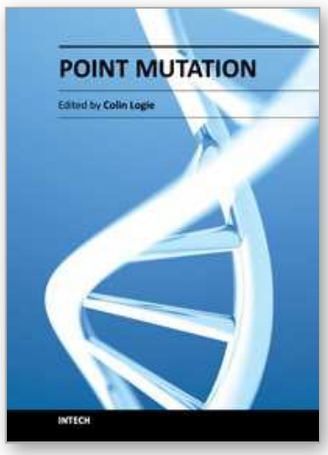
Tuch, B B; Li, H & Johnson A D (2008) Evolution of eucaryotic transcription circuits. *Science* 319:1797-1799

Wagner, A (2007) Rapid detection of positive selection in genes and genomes through variation clusters. *Genetics* 176:2451-2463

Yampolsky, L Y; Kondrashov, F A & Kondrashov, A S (2005) Distribution of the strength of selection against amino acid replacements in human proteins. *Hum.Mol. Genetics* 14:3191-3201

Zhao, Z & Zhang, F (2006) Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome. *Gene* 366:316-324

Zhang, Z & Gerstein, M (2003) Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acid Res.* 31:5338-5348



Point Mutation
Edited by Dr Colin Logie

ISBN 978-953-51-0331-8
Hard cover, 352 pages
Publisher InTech
Published online 21, March, 2012
Published in print edition March, 2012

This book concerns the signatures left behind in chromosomes by the forces that drive DNA code evolution in the form of DNA nucleotide substitutions. Since the genetic code predetermines the molecular basis of life, it could have been about any aspect of biology. As it happens, it is largely about recent adaptation of pathogens and their human host. Nine chapters are medically oriented, two are bioinformatics-oriented and one is technological, describing the state of the art in synthetic point mutagenesis. What stands out in this book is the increasing rate at which DNA data has been amassed in the course of the past decade and how knowledge in this vibrant research field is currently being translated in the medical world.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen