Genomic mapping and molecular processes

Christian Gautier, Vincent Navratil, Christelle Melo de Lima
LBBE, Universit Lyon I, 43 bd 11 novembre,69622 Villeurbanne Cedex, FRANCE

Abstract

This paper describe both management and analysis of genomic mapping data. A UML representation of both vertebrates genome maps and evolutionary relationships between gene is presented. Statistical analysis has focused on isochore organisation, subtitution rate and skew. Natural selection versus mutational bias is discussed.

Keywords:Genomic mapping, evolution

1 Introduction

Gene density, gene structure as well as genomic sequence statistical properties vary along genomes and define regions that are more or less homogeneous. These regions interact with three levels of biological constraints: i) genetic information including genes, regulatory elements, ...; ii) the management process of this information (transcription or replication units, recombination process, ...); iii) the spatial organisation of the genomic DNA with the different packaging level of chromatin (nucleosomes and higher order organisation). An important step in understanding the functioning of genomes is to associate statistical properties of sequences to each of these biological constraints. However this analysis must take into account evolutionary processes that have generated and that maintain these associations. Genomic patterns so results from a combination of several levels of constraints, natural selection and mutational bias. Inferring processes from patterns is a very complex task, this paper tries to show that taking into account spatial organisation could be of great help in genomic sequence analysis. This strategy needs developping new methodological tools both to manage and to analyse data. In this paper we will mainly focus on data base management in wich we are embedded from several years [1]. However some reference to statistical developments will be made.

Prokaryotic and eukaryotic genomes have quite different behaviors relatively to spatial patterns. We will present separatly these two groups of organisms and will focuse for eucaryotes on vertebrates.

2 Procaryotic genome patterns

Complete procaryotic genome has been the first homogeneous unit describe [2] and its discussion take always an important role in the debate between mutation bias versus selection in genome patterning. Sueka, as soon as in xxxx, determined experimentally the G+C content of bacterial genomes. It appears that those genomes show a very large range of G+C content. This raises the debate between two hypothesis: i) G+C content is linked to the fitness of the organism and its level results from a natural selection process; ii) the G+C content range inside bacteria results from a varability of the mutational bias. So the G+C content variability takes place inside the neutralist vs selectionist debate. Due to the fact that G.C link is stronger than A.T link, it has been postulated that optimal growth temperature (T_{opt})determine selection pressure acting on G+C content. Correlation studies between T_{opt} and G+C content have ruled out this hypothesis. Sueoka propose then his hypothesis of neutral modification of the replication apparatus implying variation of mutational bias. More recent studies [3] have precised relationships between temperature and G+C content. If no relation exists between genomic G+C content and T_{opt}, T_{opt} is correlated to the G+C content of genome regions coding for helix part of ribosomal RNA. This shows that probably two evo-

lutionary processes act on G+C content: mutation bias patterns the whole genome and results in the large range of bacteria genomic content, natural selection maintains a high G+C content in the region where RNA molecule structure must be conserved for high T_{opt}.

Superposition of genome patterning due to mutational bias and natural selection process is also examplifies by several structures linked to replication process. Full discussion, bibliography and continously updated analysis of complete bacterial genomes is provided by J. Lobry on the web site: http://pbil.univ-lyon1.fr/software/Oriloc/ . These data are generated by specific functions inside the R package project *SeqinR*. We will just summarize here the two main structures. It could be demonstrated that under the hypothesis of similar mutational process on each strand the equilibrium is characterized by equal amount of A and T on one side and of C and G on the other side [4]. Transforming bacterial genomes in a walk on a line define by a +1 step for C (resp. A) and -1 step for G (resp. T) put into evidence strong tendencies that delimitates for most genome two regions corresponding to the replicons. See for example Borrelia burgdorferi NC001318 or Escherichia coli CFT073 NC004431 from the Oriloc site. This approach provides an efficient estimation of the localisation of replication origin and terminus [5]. Biological interpretation of the pattern relies on the dissymetry of replication: replication works continuouly during lagging strand replication but discontinuouly when leading strand is processed (and so lagging strand is generated). The fact that this process implies that leading strand remains in a single strand state longer than the lagging strand has been particularly related to the dissymetry of mutations. This regional statistical pattern of bacterial genomes is one of the best examples where a spatial statistical analysis of genomes has led to put into evidence a biological mechanism. However another strong dissymetry can be shown between lagging and leading strands. The same type of walk along the genome with +1 when a gene is on the leading strand and -1 if it is coded on the lagging one shows also a very clear tendency (see Oriloc site). Genes are more often coded on the leading strand than on the lagging strand. This asymmetry in gene direction is considered as resulting from natural selection acting to avoid head-on collisions between replication and transcription [6].

3 Vertebrates genomes

3.1 available data

Since human genome sequencing mouse and rat genome sequence have been determined and dog genome will be known before the end of 2004. Available data on vertebrate genome is so increasing at a very great exponential rate, probably greater than the rate at which data base can be efficiently updated and sequences annotated. Moreover biological information necessary to annotate sequence is not accumulated at the same rate and if genomic sequence itself is quite well known it is far to be the same for genetic elements (genes, promoters, ...). So presently many genes are only determined through mathematical estimation and comparative analysis.

Genomic sequence is not the only means to position genetic elements on the genome and so *sequence maps* must be compared to other maps. "Cheaper" maps (particularly radiation hybrid maps "RH maps") allow comparaisons between model species for which sequence is known to species of interest (farm animals for example) for which RH map are built.

Genetic analysis of families allow to build genetic map, allowing both to position genes and to determine genetic distances and local recombination rate. It is important to note that genetic distances between markers is not a linear function of the physical distance measured by the number of nucleotides between the markers. We will see that the varaition of recombination rate along the genome could have important statistical consequences on sequences. Building genetic maps is a complex task, particularly for vetebrate having large generation timespan. A more efficient approach is provided by radiation hybrid mapping which provided distances well correlated to physical distances (RH mapping).

When two vertebrate genome maps are compared, conserved segments appears in which genes have similar orders. We will not discuss here of a precise definition of *conserved segments* keeping

as criteria a "good" order similarity. More generally the set of genes belonging to a chromosome in one species and which orthologs belong to the same chromosome of another species constitute a *synteny*. Fig. 1 provides an example of conserved segment between human chromosome 12 and mouse chromosome 6.

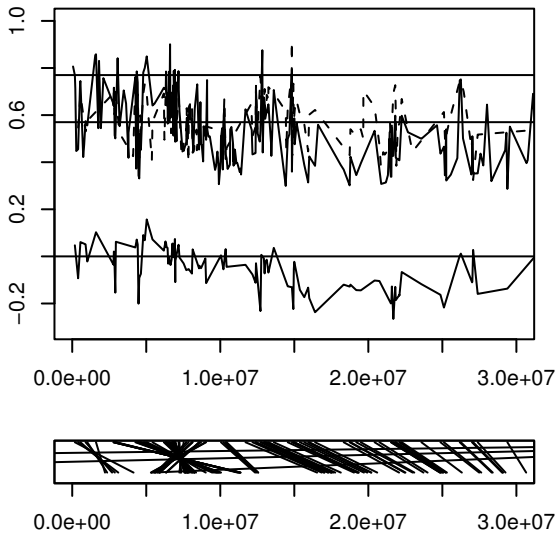

Figure 1: Synteny between human chromosome 12 and mouse chromosome 6

To relate genome map and biological processes, position of many biological informations must be compared. These informations could be statistical ones, as C+G content in third codon position of genes, but also could refers to gene expression (tissue in which the gene is expressed through EST or SAGE data) or evolutionnary process. In this last case the concept of orthology and paralogy is important. Due to duplications inside genomes sequences with great similarity could refer to different genes. More precisely orthologous genes are defined as genes in two different species such that the path linking them in the phylogenetic tree does not go through a duplication. Association of genome maps of two species is based upon association of orthologous genes. So it is of great importance to have a precise strategy to estimate orthologous pairs. Two strategies are used in litterature one is the double reciprocal best fit using blast, the second one use the analysis of the gene family phylogeny to directly verify the preceeding condition. Here we use the second approach, this can be largely automated by use of Hovergen data base [7] and the retrieval software which is able to select trees having specific patterns. The substitution process parameters (particularly K_S and K_A) can then be computed on orthologous pairs and mapped on one of the two genomes.

3.2 GemCore

GeM is a project that implies a collaboration between our laboratory and a computer scientist one (Helix, INRIA). The aim is to build a knowledge base devoted to comparative genomic mapping. In a first step [1] an UML modelling has been made as well as au GUI interface dedicated to graphical representation and request on human and mouse genome. To improve the efficiency of the software a translation in a relational date base has recently be made by V. Navratil using Postgres. This data base presently include human, mouse and rat genome data and is designed to be extended to take into account data from farm animals (pig, cow, chicken) as well as dog (for medical pupose). GeM is able to manage simultaneously all types of mapping. Moreover this new implementation include new data as:

- a direct link to sequence through the ACNUC sequence management system [8,9]
- statistical data (G+C content in each codon position, Ks between human and mouse)
- expression data (EST)
- polymorphism

Some parts of the UML modelling of GemCore is presented in the figure 2. The recursive definition of maps (due to the reflexive relation *mapsOn*) allows the inclusion of a smaller map inside a larger one, for example a local RH map inside a chromosome map. Sequence can be access through the ACNUC software (see http://pbil.univ-lyon1.fr/databases/acnuc/acnuc_gestion.html for implementation details). GenomicElement (for instance proteic genes) belongs to the set of their orthologous conterparts in the species present in

the base (OrthoGroup). These sets is then grouped inside HomolFam to represent all homologous relationships.

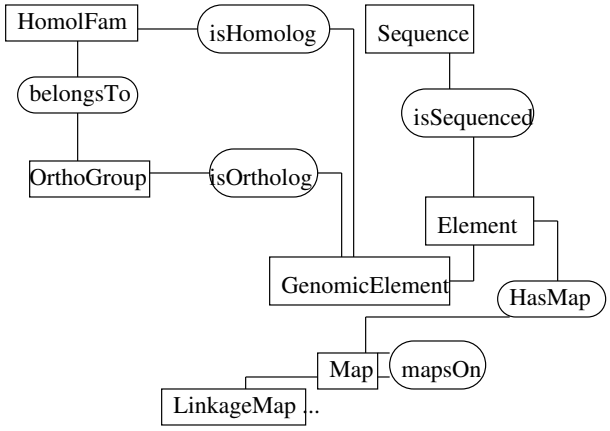

Figure 2: A part of the UML scheme of GemCore

A connection between GemCore and the R statistical software is under development using the Rdbi package. Fig. 1 results from these developments.

3.3 some regional patterns in vertebrates genomes

The great increase of available data on vertebrate genomes allows the study of the regional patterns of many genome characteristics (some related review can be founded in [10,11]. The regional pattern of nucleotide frequencies has been studied since 1976 [12] in the framework of isochore organisation. It will be presented in the next section. Substitution process is clearly involved in the generation and maintenance of such structure. The next section will summarize some results on its spatial patterns. Prokaryotic genes are organised in operon in wich gene have similar expression patterns. Works have been done to search for "operon like" eukaryotic stucture that is a clustering of genes having similar expression pattern. A short section will give some recent results from the litterature on this point. At last implications on genomic sequences of the physical structure of DNA (particularly its different folding level) has been studied since the begining of the genomic era. New data on human genome as well as the use of new methodologies have allowed new results that will be preesented in the last section.

3.3.1 isochores

The main pattern of vertebrate genome is isochore organisation which separate genome in a mosaic of regions whith a more or less homogeneous G+C content. Isochore have been discover by G. Bernardi using gradient density centrigugation [12]. Availability of genomic sequences have then allowed a much more precise description of this genome organisation. The main feature is that all type of genomic sequences is involved in isochore organisation. This results in a strong correlation between c+G content in exons, intons, intergenic regions (see for example the figure 2 of [13]). The best criteria to discriminate between isochore classes is $C+G_{III}$. In this paper we will consider that heavy (H) isochores are defined by $C+G_{III} > 77\%$ and light isochores are defined by $C+G_{III} < 57\%$. Other regions are defined as medium isochores (M). Properties of isochores are summarized here:

- density of genes is greater in H than L
- genes are larger in L than in H, this is particularly true for introns
- Lines ares more frequent in L isochores
- Sines (Alu, B1) are more frequent in H isochores

Isochores have been described mainly in mammals and birds. Figure 1 summarized some statistic features of isochores on a conserved segment between the human chromosome 12 and mouse chromosome 6. Positions of orthologous gene are link by segments in the botton of the graph. It can be clearly seen a large inversion near the telomeric end of human chromosome. The upper part of the graph is related to isochore organisation. The top solid line shows the G+C content in codons position III of human gene, the dash line is the coterpart for mouse genome. The two lines have been synchronised using interpolation between orthologous genes. The bottom line shows Δ = (human G+C_{III} - mouse G+C_{III}) for orthologous pairs. The right part show a large L isochore with negative Δ. This is characteristic of the *minor shift* : G+C content has a smaller variance in murids genome than in other mammals. In L isochores murids genome is

richer in G+C than other mammals and reciprocally murids genome has a lower G+C content than other mammals in H isochore.

Curiously the taxons where isochores was present were those in which homeothermy is known and this fact has taken an important role in proposals for the mechanisms imbedded in appearence and maintenance of isochores. Bernardi (see review in [14]) have argued for a natural selection process implying an adaptation to temperature. However the recent finding of isochores organisation in reptiles seems to ruled out the "homeothermy hypothesis" [15]. Presently no life trait or environmental factor can be related to a possible selective pressure for isochore organisation. Mutational bias provided a neutral alternative to natural selection. Two type of mutational bias have been proposed either a variation along the genome of mutational bias [16] or bias gene conversion [17]. Availability of polymorphism data is qickly increasing and seems to rules out mutational bias giving strong argument for bias gene conversion [18].

3.3.2 substitution process

It is clear that a natural selection pressure exists on nucleotides that determine the coded proteins. So it is necessary to eliminate this natural selection effect when studying relationships between substitution process and regional organisation of genomes. That may be done in taking into account only non coding region (like pseudogenes for instance) or inside coding region in using only silent substitutions. In this last case an index (K_S) of substitution has been defined to estimate the total number of substitution that have take place during evolution from the common ancester to the two considered species. Apart from some trivial processes K_S can be a property of the substitution process only if this process is stationary. A complete discussion of K_S is not in the scope of this paper and we just presented here some results:

- Matassi et al have studied the regional organisation of K_S between human and mouse [19]. This work predate the human genome sequencing and mouse genetic map have been used to localise genes. Analysis is based upon nonparametic correlation coefficient and simulations. The result was the existence of a regional organisation of K_S and its independance from the isochore organisation.

- Duret et al [13] have shown that G+C content is not at equilibrium and that the isochores organisation of mammalian genomes is vanishing.

3.3.3 expression pattern

Several studies proposes that genes having some similar expression pattern show a significant tendency to be linked in genome. So housekeeping genes in human [21], essentials genes in yeast [22] or similar expression breadth m Vanishing GC-Rich Isochores in Mammalian Genomes. ouse genes [23] made cluster in the genome. This is an important pattern suggesting that selection could apply on rearrangement to generate a suitable expression pattern. However it must be quoted that the large number of genes now available lead to statistically significant figures corresponding to very small feature. Correlation of less than 0.1 are often quoted for instance. That raised a methodological question that need to be more precicely examined.

4 conclusion

Many biological processes work on genome with a spatial component. Transcription, replication are clear examples. Their functionning imply constraints that are superimposed one on the other and that interact with genetic information. Taking into account the spatial component of observed pattern help to interpret them, procaryotic genomes give clear example of this approach.

Eucaryotic genomes undergo supplementary processes, particularly those liked to recombination. Variation of the recombination rate, existence of bias conversion raise promising hypothesis in the understanding of isochore organisation. It also imply to reinforced the link with population genetic, for example through the relationships between recombination and selection efficiency. In this context the tremendous increase of polymorphism data that will be generated particularly by EST data, can open very exciting new approaches.

In this context the availability of formal modelling of all this very different type of data and of the knowledge in the field of genomic mapping will be a crucial step to efficiently manage the tremendous data accumulation. It is the aim of the project GeM.

Moreover works must be done to develop mathematical tools capable of separate patterns having different scale. A good example is given by recent work on the analysis of sequence patterns linked to folding of eucaryotic DNA, particularly in relation with nucleosome. Wavelet analysis can focuse on some structure by choice of the used filter [24]. Markov modeling seems also to be a very promising tool to compare large regions uncluding different types of sequence (exon, intron, transposable elements, ...).

References

[1]Bronner G, Spataro B, Gautier C, Rechenmann F. (2000) GeMCore, a Knowledge Base Dedicated to Mapping Mammalian Genomes LNCS . 2066:12-23

[2] Sueoka N, Marmur J, Doty P: Heterogeneity in deoxyriboneucleic acids. II. Dependence of the density of deoxyribonucleic acids on guanine-cytosine. Nature 1959, 183:1427-1431.

[3]Galtier N, Lobry JR. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol. 1997 Jun;44(6):632-6.

[4] Lobry, JR, Lobry C (1999) Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. Mol. Biol. Evol., 16, 719-723

[5] Picardeau M, Lobry JR, Hinnebusch BJ: Physical mapping of an origin of bidirectional replication at the center of the Borrelia burgdoferi linear chromosome. Mol Microbiol 1999, 32:437-445.

[6] Brewer BJ: When polymerases collide: replication and the transciptional organization of the E. coli chromosome.(1988) Cell, 53:679-686.

[7]Duret L, Mouchiroud D, Gouy M. HOVERGEN: a database of homologous vertebrate genes.(1994) Nucleic Acids Res. 22:2360-2365

[8] Gouy M, Gautier C, Attimonelli M, Lanave C, di Paola G. ACNUC–a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. (1985) Comput Appl Biosci.1:167-172

[9] Perriere G, Combet C, Penel S, Blanchet C, Thioulouse J, Geourjon C, Grassot J, Charavay C, Gouy M, Duret L, Deleage G. Integrated databanks access and sequence/structure analysis services at the PBIL.(2003) Nucleic Acids Res.31:3393-3399

[10] Gautier C (2000) Compositional bias in DNA. Current Opinion in Genetics and Development, 10, 656-661

[11] Duret L. Evolution of synonymous codon usage in metazoans. (2002) Curr Opin Genet Dev. 12:640-649

[12] Macaya, G., Thiery, J.P., Bernardi, G. (1976) An approach to the organization of eukaryotic genomes at a macromolecular level J. Mol. Biol. 108, 237254.

[13] Duret L., Semon M., Piganeau G., Mouchiroud D., Galtier N. (2002) Vanishing GC-Rich Isochores in Mammalian Genomes. Genetics, 162, 1837-1847

[14] Bernardi G. (2000) Isochores and the evolutionary genomics of vertebrates Gene, 241, 3-17

[15] Hughes, S, Zelus D, Mouchiroud D (1999) Warm-blooded isochore structure in Nile crocodile and turtle. Mol Biol Evol 16,1521-1527

[16] Wolfe KH, Sharp PM, Li WH (1989) Nature, 337, 283-285

[17] Eyre Walker A (1993) Proc R Soc Lond B 252, 237-243

[18] Smith NG, Eyre Walker A (2001) Mol Biol Evol 18,982-996

[19] Matassi G., Sharp P.M., Gautier C. (1999) Chromosomal location effects on gene sequence evolution in mammals. Current Biology , 9, 786-790

[20]Piganeau G., Mouchiroud D., Duret L., Gautier C. (2002) Expected relationship between the silent substitution rate and the GC content: Implication for the evolution of isochores. J. Mol. Evol., 54, 129-133

[21] Lercher M.J.,Urrutia M.O., Hurst L.D. (2002) Nature Genetics, 31, 180-183

[22]Pl1 C, Hurst L.D.(2003) Evidence for coevolution of gene order and recombination rate. nature genetics, 33,392-395

[23] Hurst L.D., Williams E.J.B. (2000) Gene, 261, 107-114

[24] Audit B., Vaillant C., Arneodo A., d'Aubenton-Carafa Y., Thermes C. (2002) J. Mol. Biol., 316, 903-918