UNIVERSITAT DE BARCELONA

FACULTAT
FARMÀCIA

DEPARTAMENT
BIOQUÍMICA I BIOLOGIA MOLECULAR

LES PROPIETATS FÍSIQUES DE L'ADN EN ESCALA GENÒMICA

Josep Ramon Goñi Macià 2008

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA

DEPARTAMENT
BIOQUÍMICA I BIOLOGIA MOLECULAR

9 PUBLICACIONS

Published online January 15, 2004

354–360 *Nucleic Acids Research, 2004, Vol. 32, No. 1*
DOI: 10.1093/nar/gkh188

Triplex-forming oligonucleotide target sequences in the human genome

J. Ramon Goñi[1], Xavier de la Cruz[1,2] and Modesto Orozco[1,3,*]

[1]Molecular Modelling and Bioinformatics Unit, Institut de Recerca Biomédica, Parc Científic de Barcelona, Josep Samitier 1-5, Barcelona 08028, Spain, [2]Institució Catalana per la Recerca i Estudis Avançats (ICREA), Lluis Companys, 23, Barcelona 08028, Spain and [3]Departament de Bioquímica i Biologia Molecular, Facultat de Química, Martí i Franquès 1, Barcelona 08028, Spain

Received September 22, 2003; Revised November 17, 2003; Accepted December 3, 2003

ABSTRACT

The existence of sequences in the human genome which can be a target for triplex formation, and accordingly are candidates for anti-gene therapies, has been studied by using bioinformatics tools. It was found that the population of triplex-forming oligonucleotide target sequences (TTS) is much more abundant than that expected from simple random models. The population of TTS is large in all the genome, without major differences between chromosomes. A wide analysis along annotated regions of the genome allows us to demonstrate that the largest relative concentration of TTS is found in regulatory regions, especially in promoter zones, which suggests a tremendous potentiality for triplex strategy in the control of gene expression. The dependence of the stability and selectivity of the triplexes on the length of the TTS is also analysed using knowledge-based rules.

INTRODUCTION

The sequencing of the human genome (1,2) has opened the way for the design of new pharmacological therapies based on the inhibition of the synthesis of pathological proteins. The blocking of the synthesis of pathological proteins can be performed at least by two different mechanisms: (i) by inhibition of the translation of mRNA and (ii) by inhibition of the transcription of the corresponding gene. The first approach defines the 'anti-sense' strategy (in either its pure anti-sense and its RNA-interference versions), where oligonucleotides are used to specifically bind the target, the mRNA, blocking then the corresponding protein synthesis (3,4). The second approach defines the 'anti-gene' strategy that consists of blocking the transcription of specific genes by formation of a triple helix (5–9) at the target DNA duplex.

DNA triplexes were theoretically suggested by Pauling and Corey in 1953 (10), and probed experimentally by Rich and co-workers 4 years later (11). They are formed when a polypurine-rich DNA duplex binds a single-stranded polynucleotide [triplex-forming oligonucleotide (TFO)], through specific major groove interactions (reviewed in 5). Two types of triplexes have been described, based on the orientation of the third strand with respect to the central polypurine Watson–Crick strand: (i) parallel triplexes and (ii) anti-parallel triplexes. The first can be formed following three different motives: d(T·A-T), d(C·G-C)$^+$ (where protonated cytosine is needed in the third strand) and d(C·G-G), where Hoogsteen hydrogen bonds stabilize the interaction between the Watson–Crick (the first two bases of the triad) duplex and the third strand (Fig. 1). The second type of triplex is formed by three triads: d(T·A-A), d(C·G-G) and d(T·A-T), where the third strand makes reverse Hoogsteen pairs with the Watson–Crick duplex (Fig. 1). In general, under normal laboratory conditions, parallel triplexes are expected to be more stable than anti-parallel triplexes (12,13). However, their pH dependence [due to the presence of d(C·G-C)$^+$ triads] might limit their physiological stability.

The presence of a third strand introduces severe restrictions in the flexibility of the DNA, changing its ability to recognize specific proteins along the major groove (14,15), and accordingly, altering all the mechanisms controlling DNA function. This, and the specificity of the recognition process between the TFO and the duplex DNA explains the large number of potential applications of triplexes in the biomedical and biotechnological scenario (5–9). Thus, triplexes have been used to construct artificial restriction enzymes or to direct nuclease cutting in certain regions of the genome (6,16–17). Triplexes bound to cleaving agents have been successfully used to induce recombination in both episomal and chromosomal DNA in mammalian cells (18). It also has been reported by different authors that triplexes complexed with psoralen can be used to induce specific mutations in the genome (19–21). Furthermore, recent studies by Glazer's group (22,23) have shown that even when no chemical mutagen is added, triplex formation induces a dramatic increase in the rate of mutagenesis in the target duplex, probably as a consequence of the inability of the NER system to repair triplexes. These findings open the possibility to use triplexes as an alternative for knocking down/out specific genes (8,20).

Most of the biomedical impact of triplex technology is related to the well known ability of triplexes to inhibit mRNA synthesis in target genes, both *in vitro* (6,7,24,25) and *in vivo* (6–8,26–28). Some of the genes whose expression can be

*To whom correspondence should be addressed. Tel: +34 93 403 71 55; Fax: +34 93 403 71 57; Email: modesto@mmb.pcb.ub.es
Correspondence may also be adddressed to Xavier de la Cruz. Tel: +34 93 403 71 55; Fax: +34 93 403 71 57; Email: xavier@mmb.pcb.ub.es

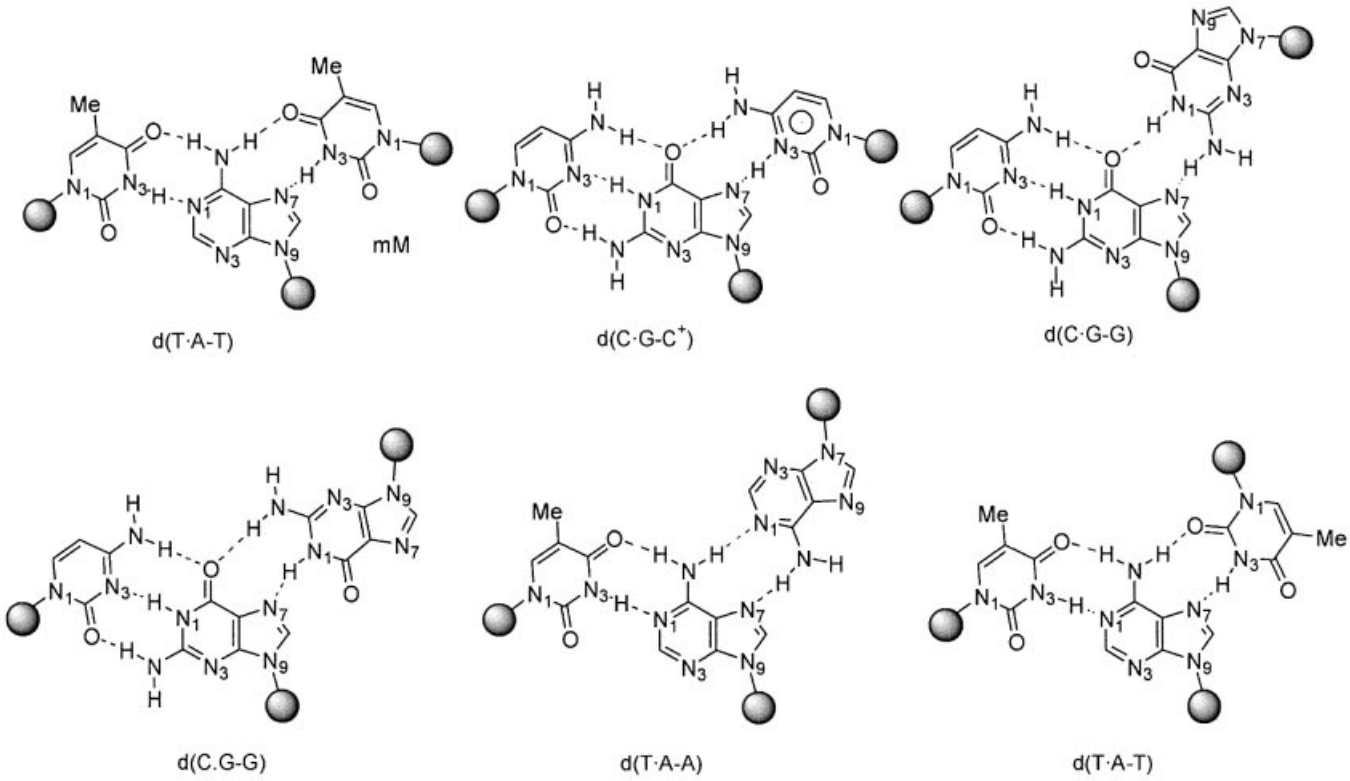

Figure 1. Schematic representation of Hoogsteen and reverse Hoogsteen-based triads in parallel and anti-parallel triplexes.

inhibited by triplex formation include genes associated with different diseases including cancer (6), suggesting that TFOs could generate, in the near future, a new generation of drugs. The inhibition of DNA transcription by triplexes can occur by means of the inhibition of mRNA elongation (29,30). However, the greatest transcription-inhibitory activity of triplexes is found when the target duplex is in the regulatory region of the gene (8).

Despite the promising results found, triplex technology still presents some shortcomings (5) mostly related to: (i) their reduced stability, (ii) sequence restrictions due to the need of polypurine tracks in the triplex target sequence, (iii) susceptibility to nucleases and (iv) problems to deliver TFOs in the cellular nucleus. A large amount of chemical, biochemical and biotechnological research is now focussed on trying to solve these practical problems of triplex technology (6,7,31,32). However, a very basic question is still unanswered: what is the triplex-forming potential of the human genome? In this paper, we perform a very extensive analysis of the human genome in order to determine how many triplex-forming oligonucleotide target sequences (TTS) exist, their location and their potential as targets for anti-gene therapy.

MATERIALS AND METHODS

Genome information

Sequence information of the human genome was taken from the UCSC database (version hg12; June 28, 2002) (http://genome.ucsc.edu/goldenPath/28jun2002) developed by the International Human Genome Mapping Consortium (1). The definition of genes, exons, coding regions, repetitive regions and conserved human–mouse regions were also taken from the UCSC Genome Browser Database (http://genome.ucsc.edu/cgi-bin/hgGateway; 33). Annotation of the genes considered in this study was obtained from the refGene (refSeq) collection, after removing redundant or overlapping genes. Promoter regions were selected as those located 100 bp upstream of the beginning of the gene. A more diffuse upstream regulatory region is selected as that located 1900 bases upstream of the promoter region. Putative downstream regulatory regions are defined 2000 bases downstream of the end of the gene. The best human/mouse conserved regions were those listed in the chrN_blatzBestMouse list in the UCSC database (http://genome.ucsc.edu/goldenPath/28jun2002). Only highly conserved blocks larger than 100 bases were considered. Repeated sequences were those obtained using the RepeatMasker software and listed in the chrN_rmsk database (http://genome.ucsc.edu/goldenPath/28jun2002; http://www.geospiza.com/products/tools/repeatmasker.htm; 34), and were used without further manipulation. Single nucleotide polymorphisms (SNPs) were mapped combining SNP databases in UCSC (snpNih and snpTsc) and the dbSNP database of genetic variation (35).

Definition of TTS

Possible TTS were defined as polypurine tracks of any size. No mismatching in the triplexes was allowed, which means that a strict triplex definition was used. In order to determine whether or not the population of TTS is that expected from a random distribution (36), we developed a simple, but flexible random model, which assuming a binomial behaviour of the

TTS distribution, allows the calculation of the expected number of TTS of a given length (in a given genome). This considers that the expected number (P) of TTS of length n in a given genome of length m can be expressed as shown in equation **1**, where q_n is the probability that a nucleotide belongs to a TTS of length n. The factor 2 appears because the TTS can happen in either DNA chains. The probability factor q_n is computed from the average number of TTS of size (n) in the random model using equations **2** and **3**. The key parameter $<i>$ is determined assuming that the number of TTS in the random model follows a binomial distribution (see equation **3**):

$$P = 2 \times m \times q_n \quad \mathbf{1}$$

$$q_n = n\ (<i> / m) \quad \mathbf{2}$$

$$<i> \approx (m - n - 1) \times \alpha^2 \times (1 - \alpha)^{n-1} \quad \mathbf{3}$$

Combining equations **2** and **3** we obtained equation **4** which provides us an approximated expression for P in a random model. Note that equation **4** is in fact an approximate solution, but provides a TTS distribution very similar to that obtained in numerical simulations of 50 randomly generated genomes with the size of the human genome (see Results). In an ideal binomial model α should be equal to 1/2 in equations **3** and **4**. However, since transitions Pur–Pyr, Pyr–Pur and Pur–Pur are not equally probable in our genome, other values of α might be more suitable. In fact, the fitting to the human genome shows that the most realistic value for α is 0.44 (the value used in the paper):

$$P \approx 2 \times n \times (m - n - 1) \times \alpha^2 \times (1 - \alpha)^{n-1} \quad \mathbf{4}$$

Prediction of triplex stability

The stability of the triplex measured in terms of the melting temperature depends on many factors, such as sequence, concentration of the TFO, length of the triplex, presence of modified nucleotides in the TFO or pH (for the most stable parallel triplexes). A rough prediction of the melting temperature of triplexes in the genome, created using the parallel motif, was determined using Roberts and Crothers empirical equations (37) (equations **5**–**7**). The stability of the corresponding anti-parallel triplexes is more difficult to determine and depends on the concentration of divalent cations, and on the possible existence of alternative structures (like the G-DNA). However, recent experiments by Eritja and co-workers (38) suggest that, in general, even for the worst cases, anti-parallel triplexes are only a few degrees less stable than the corresponding parallel triplexes at pH 4.5:

$$T_m = \frac{310 \times \Delta H^0}{\Delta H^0 - \Delta G^0_{37} - 310 \times R \times \ln\left(\frac{4}{C_{TFO}}\right)} \quad \mathbf{5}$$

where C_{TFO} is the concentration of target + triplex-forming oligonucleotides. The enthalpy (ΔH^0) and Gibbs free energy (ΔG^0) are evaluated (in kcal/mol) using equations **6** and **7**:

$$\Delta H^0 = -4.9(CC) - 8.9(TC + CT) - 7.4(TT) \quad \mathbf{6}$$

where XX means the number of dinucleotides of this particular type in the TFO:

$$\Delta G^0_{37} = -3.00(C) - 0.65(T) + 1.65(CC) + 6.0 + (C)(pH - 5.0)[1.26 - 0.08(CC)] \quad \mathbf{7}$$

where (X) means the number of nucleotides of type X in the TFO.

For discussion purposes we have considered three different TFO concentrations (μM and nM), assuming very dilute concentration for the TTS. As a reference, state of the art delivery methodologies allow the delivery of up to 20–70 μmol of TFO in the interior of the nuclei (31). Two different pH values were considered, physiological (7.0) and acidic (4.5). The use of the latter provides insight into the stability of triplexes formed with TFOs containing modified nucleotides. Triplexes with melting temperature above 50°C were considered as stable. This high temperature implies that we are using a conservative threshold of triplex stability, and probably more triplexes than those detected here can be stable under physiological conditions.

To determine the differential stability of triplexes in the human genome we compared our calculations with two background models. The first (labelled Random) assumes equal populations of A and G and equal possibility for Pur/Pur, Pur/Pyr and Pyr/Pur transitions, the second one (labelled Random H.G) was generated with the restrictions necessary to maintain the ratio A/G and the transition probabilities at the values found in the human genome. In both cases, 50 random genomes with the same length as the human genome were generated.

All the software developed here for the localization and analysis of TTS is available as C-programs upon request to the authors.

RESULTS AND DISCUSSION

The amount of nucleotides appearing in TTS in the human genome is several times larger (Fig. 2) than that expected from a random distribution (see Materials and Methods), and the difference increases as does the size of the tracks. Thus, TTS longer than 20 nt are very rare in random systems, whereas TTS longer than 30 nt are commonly found in the human genome. The existence of such a large density of TTS (or nucleotides in TTS with respect to total nucleotides) could be due in principle to two different phenomena: (i) massive duplication of a small number of TTS (this will imply a large amount of TTS in repetitive DNA) and (ii) the existence of a subtle biological effect related to these sequences. We have investigated both possibilities.

The density of TTS (defined as the number of nucleotides in TTS related to the total number of nucleotides, i.e. the probability of a nucleotide being part of a TTS) for different chromosomes is similar (data not shown), the largest density of TTS is found in the 19 chromosome and the lowest in the Y chromosome. The analysis of the base composition shows, in general, a larger population of adenines than guanines. For example, for TTS of 15 nt in length there are ~60% adenines, and for TTS of 30 nt or more the percentage of adenines

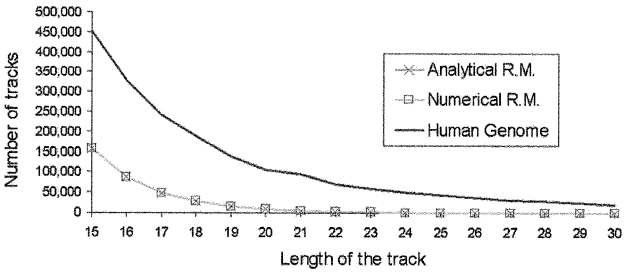

Figure 2. Number of TTS tracks of different lengths found in the human genome and in random models. The profile found in human genome is compared with that found in analytical or numerical random models (see text for details).

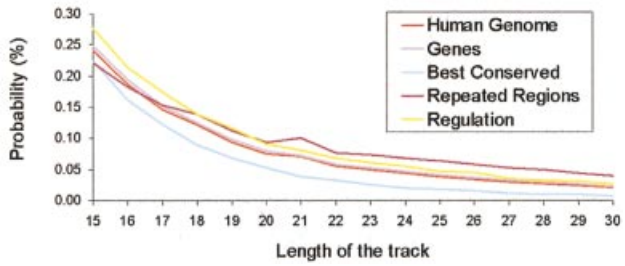

Figure 3. Probability (in %) of a nucleotide being part of TTS of different lengths in different regions of the human genome.

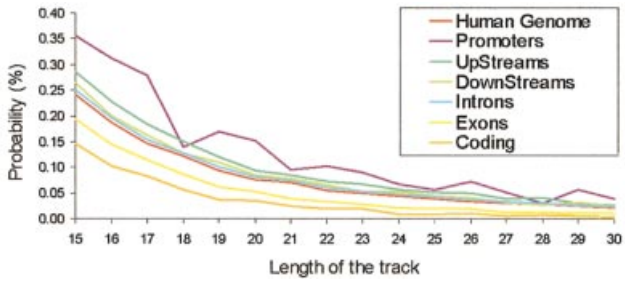

Figure 4. Probability (in %) of a nucleotide being part of TTS of different lengths in different parts of the gene region.

increase to almost 80%. It is worth noting that adenines constitute ~60% of the purines in our genome, which means that in long tracks, adenines are over-represented in TTS. We should note that the larger population of adenines in TTS is favourable from the point of view of triplex formation, since Watson–Crick guanines are targeted (Fig. 1) either by Hoogsteen cytosines (leading to a strong pH dependence of the triplex) or by reverse Hoogsteen guanines [leading then to a strong competition in the TFO between single-stranded (that are needed for triplex formation) and tetraplexes] (5).

To analyse whether or not TTS are located in regions of importance for anti-gene therapy, we divided (see Materials and Methods) the human genome into repeated regions (~50% of the genome), genes (the 10 000 of the RefGene collection), best human–mouse conserved regions (~5% of the genome) and general regulatory regions (2000 bases up and downstream of the 10 000 selected genes). The profile of TTS found for the genes (Fig. 3) reproduces very well the corresponding profile found for the total human genome. On the contrary, the highly conserved human–mouse region shows a lower ability to make triplex than the average genome, which can be partly explained by the fact that a good portion of highly conserved regions are protein-coding regions, where, as noted below, a low quantity of TTS is found. Interestingly, repeated and regulatory regions exhibit larger concentrations of long TTS than the average for the genome (Fig. 3).

In order to analyse in more detail the presence of TTS in regions of special relevance, we divided the gene region into: (i) exons, (ii) coding sequences (i.e. the exons after removing UTR), (iii) introns, (iv) promoter regions (100 nt upstream of the beginning of the gene), (v) regulatory region upstream (1900 nt) promoters and (vi) downstream (2000 nt) regulatory regions. Very interestingly (Fig. 3), only two lines appear below that of the global human genome: the exon and the coding regions. This is likely to reflect the compositional bias in the nucleic acid sequence associated with the coded polypeptide. The relative number of nucleotides in TTS is larger in all regulatory regions than in the whole genome (Fig. 4). Interestingly, the very short promoter region contains a very large concentration of nucleotides in TTS (Fig. 4), indicating that the crucial region of control of gene expression can be easily targeted for triplex formation. The existence of this large concentration of TTS in a key region of the genome strongly suggests some subtle biological function for this type of sequence, which might be related to some structural properties which can favour interaction of DNA with control proteins. In any case, for the purpose of this paper we must emphasize that the existence of a large density of TTS in the promoter region provides a tremendous opportunity for the use of triplex-based approaches in anti-gene therapies. In fact, many genes of possible therapeutic impact show extremely large TTS (>40 nt) in the promoter region. Examples are SCA1 (the gene causing type 1 spinocerebellar ataxia), ATP6V1B1 (the gene encoding for ATPase related to sensorineural deafness), PAWR (a key gene in apoptotsis), HIPK3 (a kinase involved in multidrug-resistant cells), SOX10 (mutations in this gene leads to Waardenburg–Hirschsprung disease), NOVA1 (an onconeural antigen related to breast and small cell lung cancer), and many other examples that will be discussed in more detail in a further communication.

Repetitive regions represent ~49% of the human genome (1,2), and are clearly those where sequencing is more difficult, and where a larger portion of gaps in the genome sequence exists. Our analysis shows that, as a whole, repetitive regions have a high density of TTS, just after the TTS-rich promoter region, a result that is not unexpected considering that a part of the repetitive DNA is defined by polypurine tracks. In order to study more precisely the distribution of TTS in repetitive regions, we analysed TTS density in different classes of repetitive DNAs: (i) long interspersed elements (LINE), (ii) small interspersed elements (SINE), (iii) long terminal repeat (LTR), (iv) transposons and (v) unclassified repeated DNA (low complexity, simple repeated, satellites and others). LTR and LINE regions show a density of TTS similar to that of the 'no repeated' part of the genome (Fig. 5). DNA transposons show a TTS density lower than the no-repeated part of the genome, denoting their origins as coding sequences (see

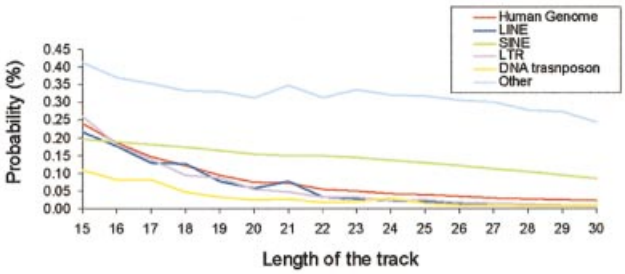

Figure 5. Probability (in %) of a nucleotide being part of TTS of different lengths in different parts of the repetitive DNA.

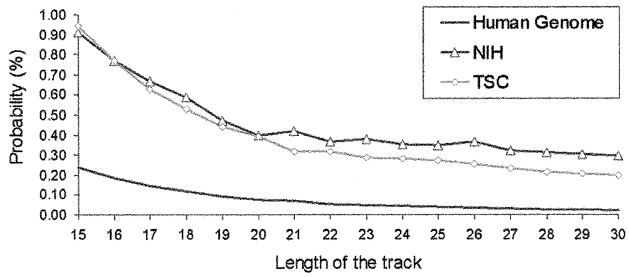

Figure 6. Probability (in %) of a nucleotide being part of TTS containing SNPs (results for both snpNih and snpTsc databases are included). The base line corresponding to the whole human genome is displayed for comparison.

discussion above). SINEs present a density of TTS larger than that of the non-repetitive part of genome, and quite interestingly, a TTS versus length of the track (in the sequence-length considered) profile quite different to the exponential decay found in the other cases (Fig. 5). This suggests that at least a part of SINEs was generated by massive duplication of a number of different short/medium sequences of DNA, biasing the distribution of possible TTS. Finally, unclassified repeated sequences also show a non-exponential decay with length. This is due to the fact that simple repeated sequences (obtained by massive replication of small DNA fragments) and low complexity DNA [rich in polypurine sequences (1,2)] are incorporated within this family of repetitive DNA.

The next step in our analysis was to follow the presence of SNPs in TTS sequences. SNPs are the major source of genetic variability in humans (39). Thus, knowing the impact of SNPs in TTS may be of biomedical interest for the purpose of designing individual-adapted therapies. Results in Figure 6 show that the probability of a position in the human genome to be part of a TTS is largely increased if it exhibits polymorphism. No obvious reason was found for this interesting behaviour other than mutations will be made more easily, or worse repaired, in TTS, or that transient triplexes are formed in polypurine tracks leading to an increase in the mutagenic rate (22,23). However, for biomedical and biotechnological purposes, what is clear is that this finding reinforces the interest of anti-gene strategies in individual-directed therapies, as taking into account the structural variability of the protein in the process of drug design is much more complex than designing oligonucleotides with varying sequences.

Previous analysis shows that TTS are more abundant in the genome than expected, and that they are particularly frequent in regions of special relevance for gene expression, suggesting *a priori* that anti-gene strategies may be very promising. However, to assess the real biomedical and biotechnological impact of triplex strategies two questions must be answered: (i) what is the selectivity TFO (i.e. what is the degree of uniqueness of a given TTS in the genome) and (ii) how stable are the triplexes formed? To answer the first question we computed how many of all the possible combinations of TTS of a given length are present in the human genome, and in the gene region (genes + regulatory regions). As shown in Figure 7, the human genome samples all possible sequence space in TTS shorter than 17, and no selectivity is possible for the complementary TFOs. As noted in Figure 7, the percentage of TTS sampled in the human genome decreases below

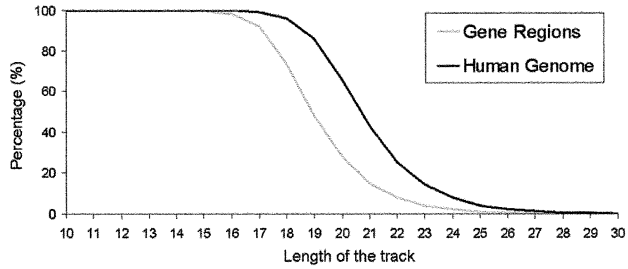

Figure 7. Coverage of the space of sequences of TTS in the human genome and in the gene region.

50% for TTS longer than 21 nt, and TTS are almost unique for lengths >26 nt, making it possible to design 100% specific TFOs. If only the gene region is considered in the analysis, the minimum length of the TTS needed to define specific TFOs is smaller [19 (one secondary interaction is expected on average) and 24 (no secondary interaction expected)]. In summary, selectivity can be reached with relatively small oligonucleotides which can be easily introduced inside the cell (6,7,31,32).

In order to determine the stability of triplexes under physiological conditions we follow Crother's empirical rules (see Materials and Methods), which allow us to determine the melting temperature of parallel triplexes based on the triplex sequence, concentration of oligonucleotides (i.e. the amount of TFO that can be internalized at the nuclei), the triplex length and the pH (acidic pH makes anti-parallel triplexes more stable). The stability of parallel triplexes largely depends on the pH. Thus, at acidic pH (4.5), ~3% of the human genome is found in TTS susceptible to form stable triplexes (melting temperature >50°C) at micromolar concentrations of TFO. When the pH is raised to 7.0 only 0.2% of the human genome can form stable triplexes. For anti-parallel triplexes, no pH dependence is expected, and on the basis of recent experiments by Eritja's group (38) the range of stabilities are expected to be those corresponding to parallel triplexes at moderately acid pH (~5–6).

The human genome shows a surprisingly good ability to form stable triplexes, much better than that expected from numerical random models (data not shown, but available upon request) at any pH, but the difference is especially large for neutral pH, indicating that triplexes in the human genome are

less sensitive to the pH than expected. The larger density of stable triplexes is found in the regulatory regions, in particular in promoter regions, where triplexes are stable even in unfavourable pH conditions (data not shown, but available upon request). In summary, our results strongly suggest that there is a large percentage of the human genome that can lead, under physiological conditions, to stable triplexes. In fact, our calculations strongly suggest that in the design of TFOs the requirement for selectivity is stricter than the requirement for stable triplexes.

CONCLUSIONS

In silico analysis of the human genome allows us to identify regions that can lead to triplex formation when a suitable TFO is available. The regions susceptible to form triplex (TTS) are more common in the human genome than expected by random models, even when these models are adapted to the composition of the human genome. A large density of TTS, which can yield stable and specific triplexes under physiological conditions, are found in promoter regions, opening interesting possibilities for the use of triplexes in the control of gene expression. Also, interestingly, TTS present an unusually large number of SNPs, which suggests that triplex strategies may be of interest in individual-oriented therapies (in fact, several examples are found of SNPs located in large TTS segments located in promoter regions). Overall, our results show the large possibilities of triplex technology in the biomedical and biotechnological scenario.

ACKNOWLEDGEMENTS

We thank Dr Roderic Guigó for many helpful discussions. This work has been supported by the Spanish Ministry of Science and Technology (PM99-0046 and BIO2003-06848). R.G. is supported by a fellowship of the IRBB-PCB.

REFERENCES

1. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
3. Hammond,S.M., Caudy,A.A. and Hannon,G.J. (2001) Post-transcriptional gene silencing by double-stranded RNA. *Nature Rev. Genet.*, **2**, 110–119.
4. Stein,C.A. and Cheng,Y.C. (1993) Antisense oligonucleotides as therapeutic agents–is the bullet really magical? *Science*, **261**, 1004–1012.
5. Robles,J., Grandas,A., Pedroso,E., Luque,F.J., Eritja,R. and Orozco,M. (2002) Nucleic acid triple helices: stability effects of nucleobase modifications. *Curr. Org. Chem.*, **6**, 1333–1368.
6. Soyfer,V.N. and Potaman,V.N. (1996) *Triple-Helical Nucleic Acids.* Springer-Verlag: New York.
7. Giovannangeli,C. and Hélène,C. (2000) Triplex technology takes off. *Nat. Biotechnol.*, **18**, 1245–1246.
8. Knauert,M.P. and Glazer,P.M. (2001) Triplex forming oligonucleotides: sequence-specific tools for gene targeting. *Hum. Mol. Genet.*, **10**, 2243–2251.
9. van Dongen,M.J.P., Doreleijers,J.F., van der Marel,G.A., van Boom,J.H., Hilbers,C.W. and Wijmenga,S.S. (1999) Structure and mechanism of formation of the H-y5 isomer of an intramolecular DNA triple helix. *Nature Struct. Biol.*, **6**, 854–859.
10. Pauling,L. and Corey,R.B. (1953) A proposed structure for the nucleic acids. *Proc. Natl Acad. Sci. USA*, **39**, 84–97.
11. Felsenfeld,G., Davis,D.R. and Rich,A. (1957) Formation of a three-stranded polynucleotide molecule. *J. Am. Chem. Soc.*, **79**, 2023–2024.
12. Scaria,P.V. and Shafer,R.H. (1996) Calorimetric analysis of triple helices targeted to the d(G3A4G3).d(C3T4C3) duplex. *Biochemistry*, **35**, 10985–10994.
13. Chandler,S.P. and Fox,K.R. (1996) Specificity of antiparallel DNA triple helix formation *Biochemistry*, **35**, 15038–15048.
14. Shields,G.A., Laughton,C.A. and Orozco,M. (1997) Molecular dynamics simulations of the d(T·A·T) triple helix *J. Am. Chem. Soc.*, **119**, 7463–7469.
15. Jiménez,E., Vaquero,A., Espinás,M.L., Soliva,R., Orozco,M., Bernués,J. and Azorin,F. (1998) The GAGA factor of *Drosophila* binds triple-stranded DNA. *J. Biol. Chem.*, **273**, 24640–24648.
16. Strobel,S.A. and Dervan,P.B. (1992) Triple helix-mediated single-site enzymatic cleavage of megabase genomic DNA. *Methods Enzymol.*, **216**, 309–321.
17. Zain,R., Marchand,C., Sun,J., Nguyen,C.H., Bisagni,E., Garestier,T. and Hélène,C. (1999) Design of a triple-helix-specific cleaving reagent. *Chem. Biol.*, **6**, 771–777.
18. Luo,Z., Macris,M.A., Faruqi,A.F. and Glazer,P.M. (2000) High-frequency intrachromosomal gene conversion induced by triplex-forming oligonucleotides microinjected into mouse cells. *Proc. Natl Acad. Sci. USA*, **97**, 9003–9008.
19. Havre,P.A., Gunther,E.J., Gasparro,F.P. and Glazer,P.M. (1993) Targeted mutagenesis of DNA using triple helix-forming oligonucleotides linked to psoralen. *Proc. Natl Acad. Sci. USA*, **90**, 7879–7883.
20. Majumdar,A., Khorlin,A., Dyatkina,N., Lin,F.L., Powell,J., Liu,J., Fei,Z., Khripine,Y., Watanabe,K.A., George,J., Glazer,P.M. and Seidman,M.M. (1998) Targeted gene knockout mediated by triple helix forming oligonucleotides. *Nature Genet.*, **20**, 212–214.
21. Barre,F.X., Ait-Si-Ali,S., Giovannangeli,C., Luis,R., Robin,P., Pritchard,L.L., Hélène,C. and Harel-Bellan,A. (2000) Unambiguous demonstration of triple-helix-directed gene modification. *Proc. Natl Acad. Sci. USA*, **97**, 3084–3088.
22. Wang,G., Seidman,M.M. and Glazer.P.M. (1996) Mutagenesis in mammalian cells induced by triple helix formation and transcription-coupled repair. *Science*, **271**, 802–805.
23. Vasquez,K.M., Narayanan,L. and Glazer,P.M. (2000) Specific mutations induced by triplex-forming oligonucleotides in mice. *Science*, **290**, 530–533.
24. Duval-Valentin,G., Thuong,N.T. and Hélène,C. (1992) Specific inhibition of transcription by triple helix-forming oligonucleotides. *Proc. Natl Acad. Sci. USA*, **89**, 504–508.
25. Cooney,M., Czernuszewicz,G., Postel,E.H., Flint,S.J. and Hogan,M.E. (1988) Site-specific oligonucleotide binding represses transcription of the human c-myc gene *in vitro*. *Science*, **241**, 456–459.
26. Grigoriev,M., Praseuth,D., Robin,P., Hemar,A. and Saison-Behmoaras,T. (1992) A triple helix-forming oligonucleotide-intercalator conjugate acts as a transcriptional repressor via inhibition of NF kappa B binding to interleukin-2 receptor alpha-regulatory sequence. *J. Biol. Chem.*, **267**, 3389–3395.
27. Joseph,J., Kandala,J.C., Veerapanane,D., Weber,K.T. and Guntaka,R.V. (1997) Antiparallel polypurine phosphorothioate oligonucleotides form stable triplexes with the rat alpha1(I) collagen gene promoter and inhibit transcription in cultured rat fibroblasts. *Nucleic Acids Res.*, **25**, 2182–2188.
28. Postel,E.H., Flint,S.J., Kessler,D.J. and Hogan,M.E. (1991) Evidence that a triplex-forming oligodeoxyribonucleotide binds to the c-myc promoter in HeLa cells, thereby reducing c-myc mRNA levels. *Proc. Natl Acad. Sci. USA*, **88**, 8227–8231.
29. Young,S.L., Krawczyk,S.H., Matteucci,M.D. and Toole,J.J. (1991) Triple helix formation inhibits transcription elongation *in vitro*. *Proc. Natl Acad. Sci. USA*, **88**, 10023–10026.
30. Faria,M., Wood,C.D., Perrouault,L., Nelson,J.S., Winter,A., White,M.R., Hélène,C. and Giovannangeli,C. (2000) Targeted inhibition of transcription elongation in cells mediated by triplex-forming oligonucleotides. *Proc. Natl Acad. Sci. USA*, **97**, 3862–3867.
31. Ebbinghaus,S.W., Vigneswaran,N., Miller,C.R., Chee-Awai,R.A., Mayfield,C.A., Curiel,D.T. and Miller,D.M. (1996) Efficient delivery of triplex forming oligonucleotides to tumor cells by adenovirus-polylysine complexes. *Gene Ther.*, **3**, 287–297.

32. Zendegui,J.G., Vasquez,K.M., Tinsley,J.H., Kessler,D.J. and Hogan,M.E. (1992) *In vivo* stability and kinetics of absorption and disposition of 3′ phosphopropyl amine oligonucleotides. *Nucleic Acids Res.*, **20**, 307–314.
33. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J., Weber,R.J. and Kent,W.J. (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
34. Smith,A.F. (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes *Curr. Opin. Genet. Dev.*, **9**, 657–663.
35. Sherry,S.-T., Ward,M.-H., Kholodov,J., Baker,L., Pham,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
36. Ussery,D., Soumpasis,D.M., Brunak,S., Staerfeldt,H.H., Worning,P. and Krog,A. (2002) Bias of purine stretches in sequenced chromosomes. *Comput. Chem.*, **26**, 531–541.
37. Roberts,R.W. and Crothers,D.M. (1996) Prediction of the stability of DNA triplexes. *Proc. Natl Acad. Sci. USA*, **93**, 4320–4325.
38. Jaumot,J., Eritja,R., Tauler,R. and Gargallo,R. (2003) Resolution of parallel and antiparallel oligonucleotide triple helices formation and melting processes by means of multivariate curve resolution. *J. Biomol. Struct. Dyn.*, **21**, 267–278.
39. Collins,F.S., Brooks,L.D. and Chakravarti,A. (1998) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.*, **8**, 1229–1231.

BMC Genomics

BioMed Central

Research article

Open Access

Exploring the reasons for the large density of triplex-forming oligonucleotide target sequences in the human regulatory regions

Josep Ramon Goñi[1], Juan Manuel Vaquerizas[2], Joaquin Dopazo[2,3] and Modesto Orozco*[1,4,5]

Address: [1]Molecular Modeling and Bioinformatics Unit. Institut de Recerca Biomèdica. Parc Científic de Barcelona. Josep Samitier 1-5. Barcelona 08028. Spain, [2]Department of Bioinformatics, Centro de Investigación Príncipe Felipe, Avda. Autopista del Saler 16, Valencia, 46013, Spain, [3]Functional Genomics Node, Instituto Nacional de Bioinfomatica, Centro de Investigación Príncipe Felipe, Avda. Autopista del Saler 16, Valencia 46013, Spain, [4]Departament de Bioquímica i Biología Molecular. Facultat de Química. Universitat de Barcelona. Martí i Franquès 1. Barcelona 08028. Spain and [5]Protein Structure and Modeling Node. Instituto Nacional de Bioinfomàtica. Genoma España. Parc Científic de Barcelona. Josep Samitier 1-5. Barcelona 08028. Spain

Email: Josep Ramon Goñi - rgoni@mmb.pcb.ub.es; Juan Manuel Vaquerizas - jvaquerizas@ochoa.fib.es; Joaquin Dopazo - jdopazo@ochoa.fib.es; Modesto Orozco* - modesto@mmb.pcb.ub.es

* Corresponding author

Published: 27 March 2006

*BMC Genomics*2006, **7**:63 doi:10.1186/1471-2164-7-63

Received: 21 December 2005
Accepted: 27 March 2006

This article is available from: http://www.biomedcentral.com/1471-2164/7/63

Abstract

Background: DNA duplex sequences that can be targets for triplex formation are highly over-represented in the human genome, especially in regulatory regions.

Results: Here we studied using bioinformatics tools several properties of triplex target sequences in an attempt to determine those that make these sequences so special in the genome.

Conclusion: Our results strongly suggest that the unique physical properties of these sequences make them particularly suitable as "separators" between protein-recognition sites in the promoter region.

Background

DNA triplexes [1] are formed when a duplex containing a poly-purine track is recognized by single-stranded poly-nucleotide (noted as the triplex-forming oligonucleotide; TFO). The third strand interacts through the major groove of the duplex, thereby making specific hydrogen bond interactions with the Watson-Crick purines [2,3]. The TFO can be DNA, RNA or different oligonucleotides with modifications in either their nucleobases or phosphoribose backbone [4]. Two types of triplexes have been described on the basis of the orientation of the TFO with respect to the central polypurine track: i) parallel triplexes and ii) anti-parallel triplexes. The parallel triplex is characterized by Hoogsteen hydrogen bonds between the TFO (typically pyrimidine-rich) and the central Watson-Crick purine [see Figure 1], while the anti-parallel triplexes show reverse-Hoogsteen hydrogen bonds and the TFO is purine-rich [see Figure 1]. Parallel triplexes are believed to be more stable than the anti-parallel ones in normal laboratory conditions, but the situation can reverse in physiological environments, especially when the target duplexes contain a poly-G track [2-5].

The presence of a TFO in the major groove of the duplex leads to major distortions in the capacity of the target duplex to be recognized by specific proteins [2,6,7]. This produces major changes in the functionality of the target duplex, which has been used for biotechnological and

biomedical purposes [2,3,8-10]. Thus, modified TFOs containing suitable chemical compounds have been used to develop artificial nucleases [11,12], to induce recombination in mammalian cells [13], and to trigger mutations in target DNA [13-16]. In all these cases, the formation of the triplex guides the active chemical compound to the proper position in the target genome. Unmodified TFOs increase the rate of mutations at the triplex target sequence (TTS), which opens the possibility for knocking down genes [9,16,17]. Triplex formation inhibits mRNA synthesis [2,8,9,18-23] when the TTS is located at a regulatory region. Furthermore, when the triplex is formed in the middle of a gene, mRNA elongation is stopped just before the TTS, which indicates that triplex binding is strong enough to displace complex transcriptional machinery [24,25]. These two findings open up the possibility to use TFOs as "anti-gene" drugs. These pharmacological agents would have the capacity to specifically arrest the transcription of pathological genes, thus leading to an intense and targeted therapeutic action [3,8-10]. However, despite their promise, anti-gene therapies still face many technical problems [2,3] and the density and location of TTSs in human genes is unclear.

In a recent paper, we explored the presence of TTSs (polypurine tracts which are expected to lead to stable triplexes in physiological conditions) in the human genome [26]. Our analysis showed that these sequences are vastly over-represented when compared to what randomness predicts. Interestingly, the largest relative concentration of TTSs occurs in the upstream regulatory region (especially at the proximal promoter region: 100 nts upstream) [26]. Recent studies by our group (Goñi *et al*. Unpublished results) show that these trends are common to many other organisms, from mammals to procaryotes, indicating that many genes may be targets for triplex formation (Goñi *et al*. Unpublished results). However, this interesting finding raises an intriguing question: why are TTSs so abundant in crucial regions for the control of genome function?

Here we present an extensive descriptive analysis of TTSs in the human genome in an attempt to elucidate why these sequences are so abundant in regulatory regions. Our results indicate that the unique physical properties of TTSs may explain this overpopulation.

Results and discussion

As proposed in previous studies using an older genomic data base [26], TTSs are largely over-represented in human

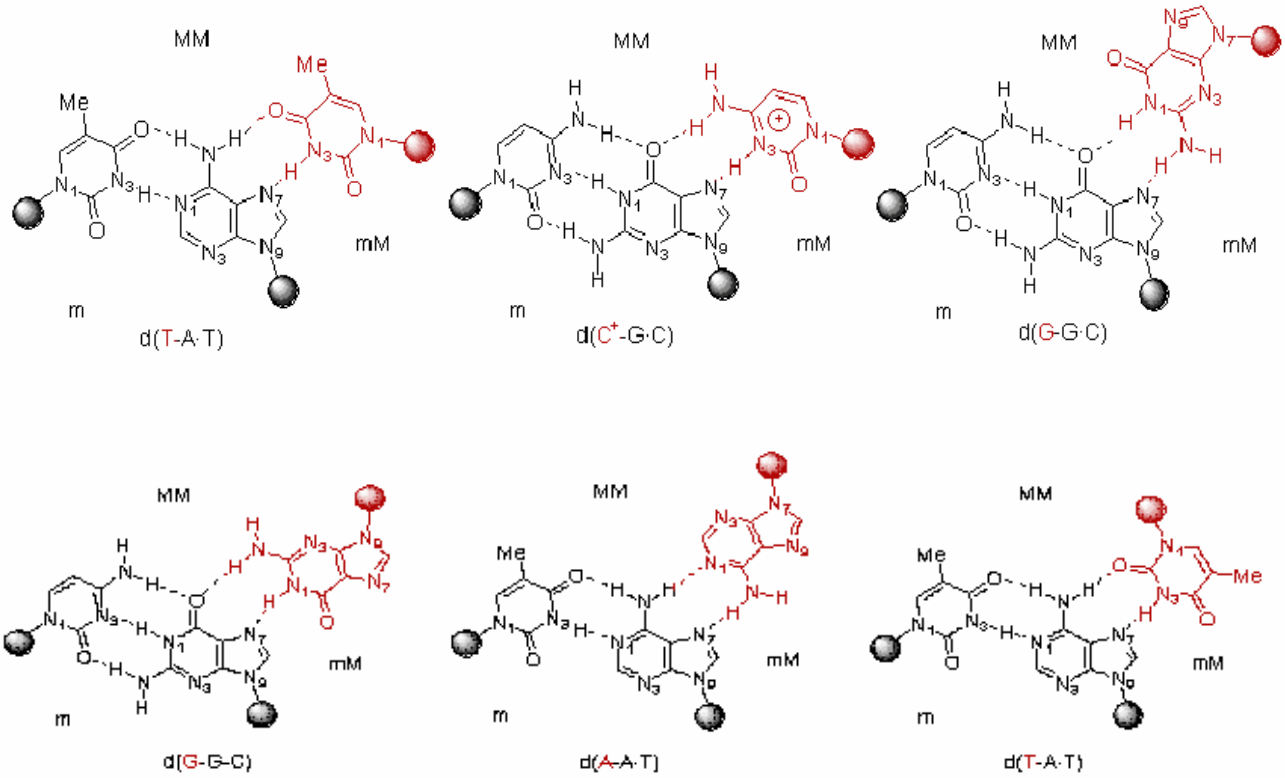

Figure 1
Schematic representation of parallel and anti-parallel triads present in triplexes.

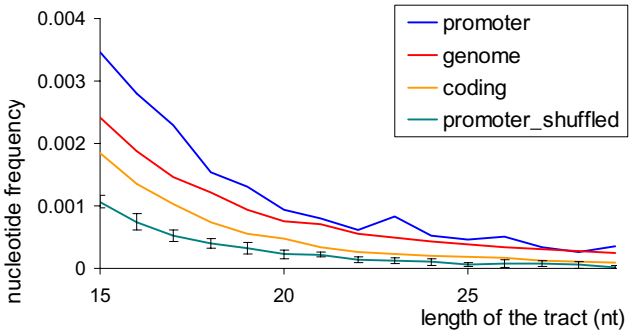

Figure 2
Frequencies of nucleotides forming part of TTSs for different lengths in the human genome, coding and promoter regions (100 nucleotides upstream). Values represented here correspond to total population from the human genome and accordingly do not have error bars associated. Random promoter expected values and its deviation is computed generating 10 sets of sequences shuffling our promoter collection.

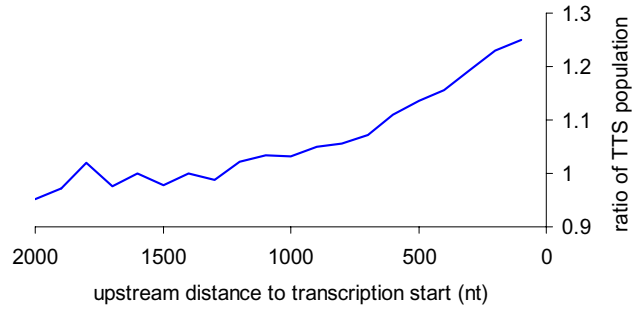

Figure 3
Ratio of TTSs (lengths from 10 to 25 nts) in promoter regions (from 2000 to 100 nts upstream) with respect to average TTS population in the human genome. Once again data is taken for the entire genome and correspond to absolute values, without associated errors.

genome with respect to a background model such as that described in reference [[26]; see Figure 2]. This over-representation is particularly noticeable when considering the proximal promoter region (100 nts upstream), where a considerable density of large TTSs are found. Note that the over-representation is clear irrespective of the random model used (for the shake of simplicity only few random models are shown in Figure 2; the rest are displayed in supplementary material [see Additional file 1]) and that the statistical significance of the difference is demonstrated by Clover calculations ($p < 10^{-20}$; [see *Methods*]).

Very interestingly, the over-representation of TTSs in the promoter region with respect to the general human genome decreases as more distant promoter regions are considered [see Figure 3]. Clearly, there is an unusually large region with potential to form triplexes in region proximal to transcription origin, which is rich in promoter regions. At first glance, several reasons for this behaviour can be offered: i) triplex formation may be an ancient regulatory mechanism [see Figure 4] for RNA-mediated control of gene expression, ii) target sequences for transcription factors have an overpopulation of TTSs, iii) TTSs have several intrinsic physical properties that are useful for protein binding to DNA.

Are TTSs part of an ancient DNA auto-regulatory mechanism?

Triplex formation is a powerful mechanism by which to modulate gene function, and, when formed in the promoter region, triplexes can knock-out or knock-down a gene. Accordingly, it is feasible that the large number of TTSs at a promoter region is related to regulatory processes. Interestingly, analysis of GO terms using the FATIGO [see Methods] program shows that the set of genes with large TTSs (15–20 nts) at promoter regions correspond to a subset of genes which differs significantly (even for the very strict Benjamin-Yekutieli adjusted p-value; [see Methods]) from the background. Irrespectively of the length of the TTS (from 15 to 25 nts) and the section of the early promoter region 0–100 or 0–200 nts upstream, genes with TTS_p are over-enriched with functions in the regulation of physiological processes, and very often are characterized as transcription factors or related protein [see Figure 5]. In fact, TTS in promoters seems to be as determinant of the functionality of genes as the CpG islands [see Additional file 2]. On the basis of this observation, we therefore propose that the presence of large TTS in the promoter region of these genes might provide advantage for the control of the expression of these genes. Furthermore, it is tempting to consider the existence of an RNA-mediated feed-back mechanism [see Figure 4] which controls the expression of the regulatory genes by triplex formation between the TTS at the promoter region of these genes and the TFO present in the intron of the regulated genes. Unfortunately, GO analysis and inspection of TRANSFAC 8.3 [see Methods] failed to provide evidence connecting transcription factors with TTS at the promoter region with those with the corresponding TFO in introns. In addition, we analysed the cases in which a transcription factor with a TTS in its promoter interacted with a second transcription factor, in order to examine whether the introns of genes regulated by this second factor contained TFOs complementary to the TTSs in the first transcription factor. However, again we did not find any relationship between genes with TTS in the promoter and those with intronic TFOs. Given the low number of cases for which experimental evidence of regulation by transcription factors is available, this nega-

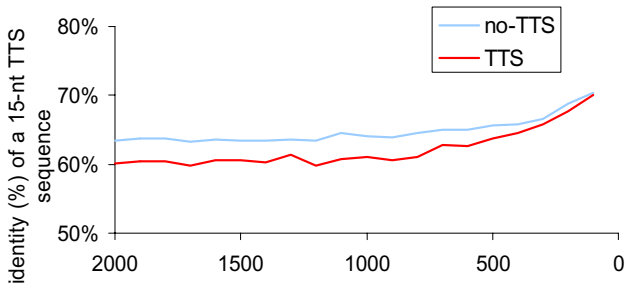

Figure 7
Percentage of human-mouse identity of 15-nt fragments in several promoter regions for TTS and non-TTS segments of the same size in equivalent regions of regulatory region.

tive result cannot be taken as evidence against the RNA-mediated feed-back mechanism proposed.

Are TTSs rich in transcription factor recognition sites?
As described in *Methods*, we mapped the TRANSFAC database into the human promoter region (up to 200 nts upstream of transcription origin) and computed the occurrences of nucleotides in long TTSs (length equal or greater than 10) in the promoter region around the transcription factor binding site (TFBS). Sequences which were recognized as targets of transcription factors showed much less probability to be in TTS than neighbouring promoter sequences [see Figure 6]. Furthermore, generation of random sequences (TTS and no-TTS) showed that no-TTS random sequences have a much larger probability to be transcription factor binding sites than TTS random sequences. Overall, we must conclude that even in some cases small (4–6) poly-purine segments might be found in TFBS, TTS as defined here (length equal or greater than 10) are very rarely present in TFBS. That means that the hypothesis that TTSs are over-represented in promoter regions because they contain TFBS can be ruled out.

Although TTSs do not interact directly with transcription factors, they show a profile of conservation when approaching to the start of transcription similar to that of the whole of non-TTS and of sequences that have been annotated as TFBS [see Figure 7 and Additional file 3]. TTS_p in the near promoter regions are quite well conserved (even not as conserved as TFBS; [see Additional file 3]) suggesting that they may have an important physiological role that is not related to direct DNA-protein interactions.

Do TTSs have distinct intrinsic physical properties?
Previous results seem contradictory and somehow difficult to rationalize. Thus, although TTS in promoters (TTS_p) are over-represented, appear in key genes for the control of physiological processes and are very well conserved, they are not targets for transcription factors. No evidence is found regarding the possibility that TTS acted as an ancient regulatory mechanism, mimicking the functionality of interference RNAs. How can we reconcile all this findings? In our view, the only possibility will be if TTS have some intrinsic physicochemical properties that are useful when present in the promoter region of certain genes.

As described in *Methods*, we analyzed several physical-chemical descriptors of DNA in two sets: i) randomly generated TTSs and ii) randomly generated human-like DNA. In general, TTSs displayed average melting temperatures similar to those of normal DNA sequences [see Figure 8], which agrees with the observation that the average stacking energy of TTSs and normal DNA sequences are the same. Thus, TTSs do not introduce bias in the stability of the DNA duplex, which could provide an advantage for the functionality of promoter regions. Curvature analysis using Bolshoy's algorithm indicates that TTSs are significantly more curved than random DNA sequences. Furthermore, analysis of configurational volume [see *Methods*] strongly suggests that TTSs are on average more rigid than normal DNA. These findings strongly support the hypothesis that TTSs at promoter regions can be used as rigid and curved separation signals for transcription factors. It is clear that these physical properties modulate nucleosome positioning and rotational phasing, and several authors have pointed out that polypurine tracks are not well incorporated into the nucleosome [27]. Unfortunately, sequence rules for nucleosome positioning and phasing are, in our hands, not accurate enough to test this hypothesis.

In addition to the possible role of TTSs in the organization of DNA in nucleosomes, when these sequences are present, their unique physicochemical properties have a large impact on the promoter region. Thus, large flexibility is desirable for DNAs that need to bind to proteins, and accordingly deform its structure, but rigidity is useful for the definition of spacing elements that should isolate protein-induced DNA deformability in specific regions of the duplex. The larger curvature is also a desired element, since it can help in the relative positioning of transcription factors in 3-D space, helping then to establish physiologically critical protein-protein interactions. Thus, the presence of TTSs at promoter regions can provide the cell with specific mechanisms, probably in most cases not related to triplex formation, by which to enhance activation/repression of genes that are crucial for the regulation of cellular processes mechanisms.

The results presented in this paper show that, irrespectively of whether the cell now uses or once used a triplex-mediated control mechanism, TTSs in the promoter region are very abundant and that genes with TTS_p are cru-

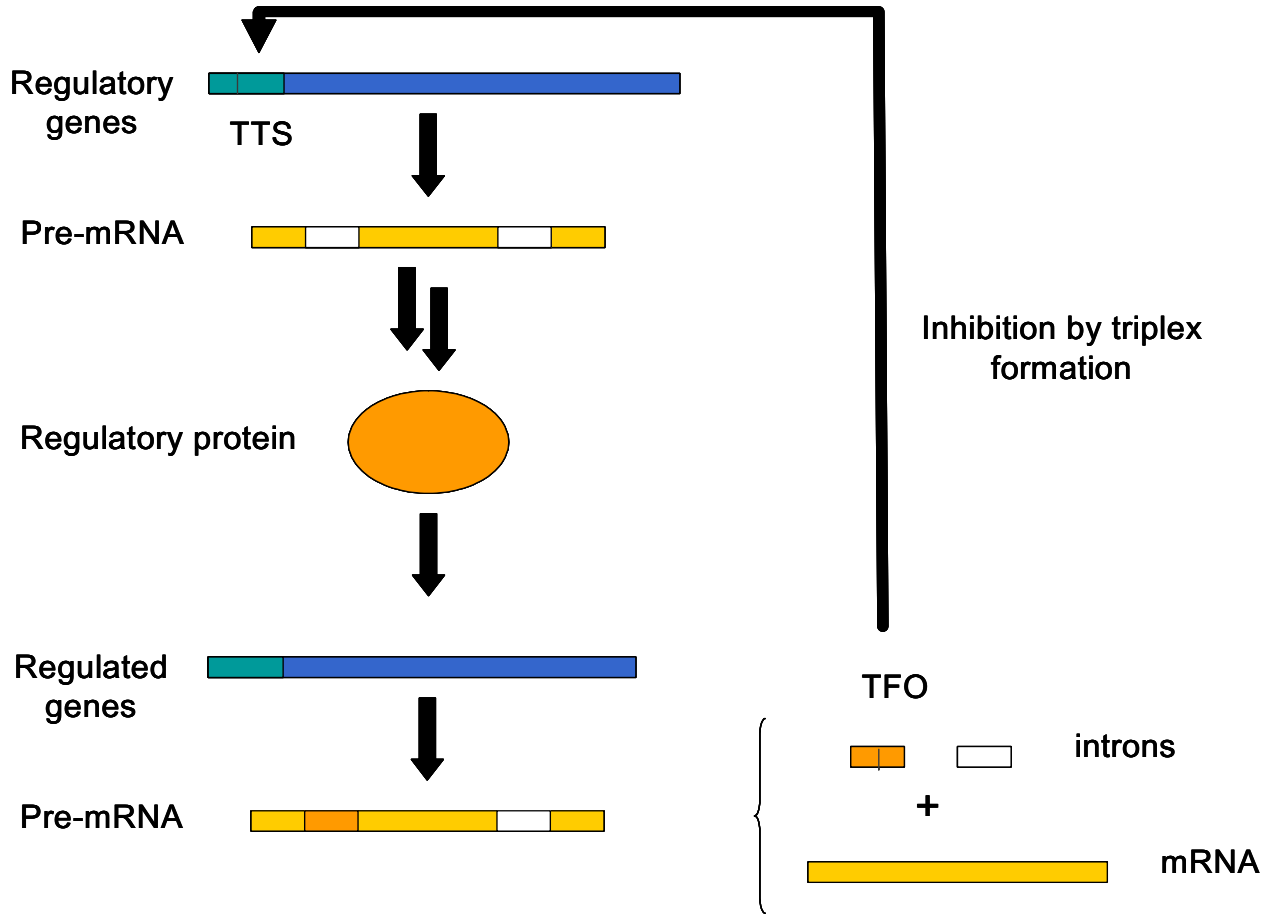

Figure 4
Putative feed-back regulatory mechanism for the control of gene function on the basis of the inhibition of regulatory genes by triplex formation between TTSs in the promoter of regulatory genes and the TFOs in the introns of regulated genes.

cial for the control of cell life. TTSp do not bind to transcription factors, but besides this, they have a conservation profile similar to that of non-TTS segments in promoter regions, including that of sequences recognizing transcription factors. Analysis reported here suggests that the TTS_p provide the promoter region unique physical properties that can contribute to a better functioning of regulatory proteins. All these results strongly support the notion that triplex-based anti-gene technology is widely applicable in the control of pathologies related to malfunctioning of the regulatory mechanisms of physiological processes.

Conclusion

Triplex-target sequences (TTS) are over-represented in the human genome. Such an over-representation is especially large when promoter regions 100 to 200 upstream are considered.

Genes with TTS in promoters are over-enriched with functions in the regulation of physiological processes, and very often are characterized as transcription factors or related protein.

TTS are not part of sequences which are directly targeted by transcription factors, but their (human-mouse) conservation profiles suggest that they are important for gene functionality.

TTS are significantly more curved and rigid than normal DNA, which suggests that (in addition to other possible functions) TTS act as spacing fragments which help in the correct positioning of transcription factors.

Methods

Genome information

Sequence information of the human genome was taken from the UCSC database [28] version hg17; May 2004;

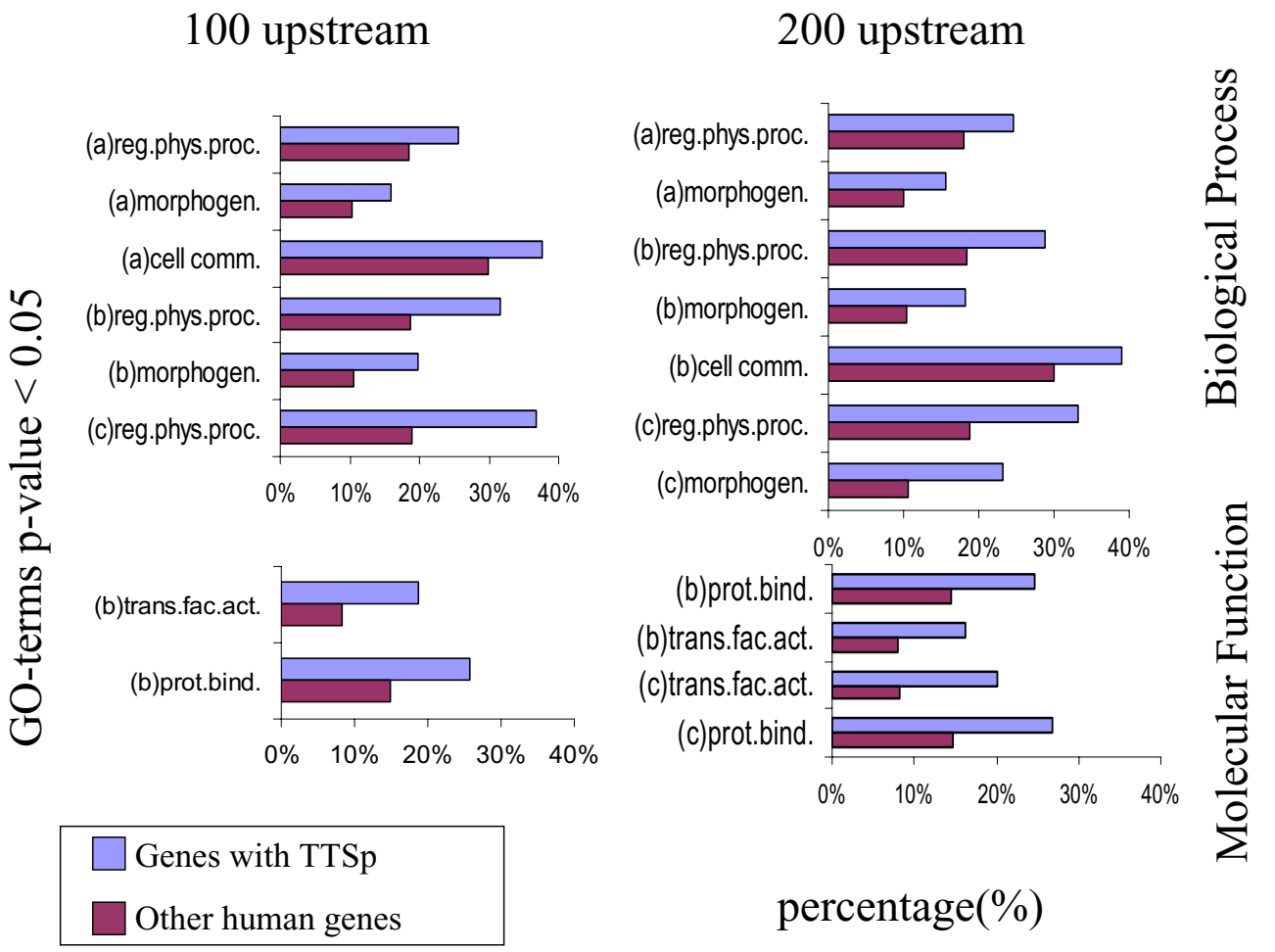

Figure 5
Results of GO analysis of biological processes (upper panels) and molecular functions (bottom panels) of genes with TTS at promoter regions. Analyses were repeated for promoter regions defined by 100 or 200 nts upstream, and considering several TTS lengths (a)15, b)20 and c)25 nts). We show only cases where i) the population is greater than 10% in the set of genes with TTS_p and ii) the subset of genes with TTS_p is significantly different $p < 0.05$ to the background genes using a very strict FDR-adjusted p-test [see *Methods*]. For all the cases shown, the normal p-test is 10^{-4} to 10^{-5}.

developed by the International Human Genome Mapping Consortium [29]. Annotation of the genes, introns, exons and coding regions considered in this study was obtained from the *refGene* (*refSeq*) collection. To avoid compute multiple times the same locus of the genome, overlapping entries of *refSeq* have been ignored. The set of upstream regions at starting positions 5000, 2000, 1000, 500, 200 and 100 nts upstream of the transcription start of *refSeq* genes were extracted from the *upstream5000* file of the same data base. The 0–5000 and specially the 0–100 regions are expected to be largely enriched in promoter sequences. USCS also provide the annotation of CpG islands (*CpGisland* file). CpG promoters are those with an overlapping feature on the 200-nt upstream region.

Definition of TTS

Possible triplex target sequences (TTSs) were defined as polypurine tracks of any size and in any strand. No mismatching in the triplexes was allowed, implying that a strict triplex definition was used. The number of TTSs would increase substantially if 5% or 10% mismatching were allowed.

Background models of TTS distributions

In order to determine the significance of a given TTS distribution we need to create background distributions. We used first an analytical binomial model [26] where all base dimmers have the same probability. To generate a more reliable background model we modified the

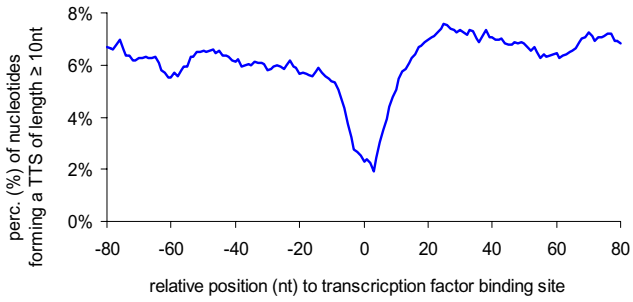

Figure 6
Percentage of nucleotides forming TTS (length equal or greater than 10 nts) at a range of distances from the centre of a transcription factor recognition site. Negative values imply upstream direction and positive values approach the transcription starting point. Calculations were performed considering only the 200 nt early promoter region.

method to account for the dimmer-distributions in human genome and also on human promoter (where Pur-Pyr, Pur-Pur and Pyt-Pur are not equally probable). We consider also a numerical background model (that at the dimmer level fits to the binomial model; [see Figure 2 in reference 26]) which allows us to introduce also trimer-biased in the promoter region. This numerical model was build by creating a 10^8-mer sequence selected which maintain the trimer (or dimer) population found in human promoters. Finally, a last random model (for promoter region) was created by using promoter-specific sufflesequ models created with the EMOSS package version 3.0 [30]. In the later case we generate 10 sets of sequences shuffling our promoter collections. A simple visual inspection shows that the real and background distributions are very different, but in any case, we confirm this by running *Clover* [31], a tool for detection of functional DNA motifs via statistical over-representation. For this purpose we create a matrix (length 15-nt), where for all positions A,G scored 1 and T,C score 0 (clover automatically scans both strands).

TTSs in promoters conserved in human and mouse

To evaluate the conservation of TTSs in the promoter (TTS_p) region, we took *upstream5000* file to build 33 mouse assemblies [28]. In order to match human and mouse promoter regions, we translated gene code to protein name using the *loc2ref* file from the NCBI database [32] for both human and mouse genes. We then searched for correspondences in *HomoloGene* Build 39.1 from the NCBI database. This procedure generated a list of 5000 pairs of human-mouse genes.

We calculated human-mouse identity for a chosen 15-nt sequence (TTS or not) from a human region, aligning it across the corresponding mouse region. The alignment was done using a bit-vector alignment algorithm [33]. For each entry, the greatest percentage of matching bases in the best alignment was processed. We estimated the conservation by averaging this value for all region entries.

For the shake of completeness the comparison was also performed using human/mouse aligned sequences present in the UCSC multiz8way (8 vertebrates) multiple alignments. As before conservation is computed by analyzing identity conservation in 15-mer sequences, averaging the data for all the 15-mer windows in the studied segment. Results obtained with this or the previous alignment protocol are very similar, reinforcing the quality of our results. Using these alignments we computed also the conservation in promoter regions annotated as transcription factor binding sites [34,35].

Functional annotation of groups of genes

To test whether a group of genes was significantly enriched in one or more functional terms (out of several thousands) with respect to the background (usually the rest of genes), we used the FatiGO algorithm [36] from the Babelomics suite [37] for functional annotation of sets of genes. This algorithm uses known functional annotations for genes obtained from the Gene Ontology (GO) consortium databases [38]. Both lists of genes (the group of interest and the background) were converted into two lists of GO terms using the corresponding gene-GO association table. For each GO term the data are represented as a 2 × 2 contingency table with rows representing presence/absence of the GO term, and each column representing each of the two lists. A Fisher's exact test for 2 × 2 contingency tables was used. Since thousands of GO terms are simultaneously tested without an *a priori* hypothesis on any particular term, p-values must be corrected for multiple testing. For this correction, we used the strict false discovery rate (FDR) method described by Benjamini and Yekutieli [39],

Hypothetical human auto-regulated TTSs

Hypothetical auto-regulated TTSs are those that appear in transcribed and promoter (200 nts upstream) regions of two distinct genes. The TFO able to recognize a TTSp is defined as the sequence that matchs with i) the complement of the TTS or ii) the reverse of the TTS. First case maps potential endogenous TFO for parallel triplex whether the second one maps the potential antiparallel TFOs [see Figure 1]. Genes were divided in three sets: i) containing TTS in the promoter region, ii) containing the corresponding TFO in the transcribed region and iii) other genes (used as background). We ran FATIGO [36,37] to identify possible relationships between genes containing TTS in the promoter and those containing the corresponding TFO in transcribed regions.

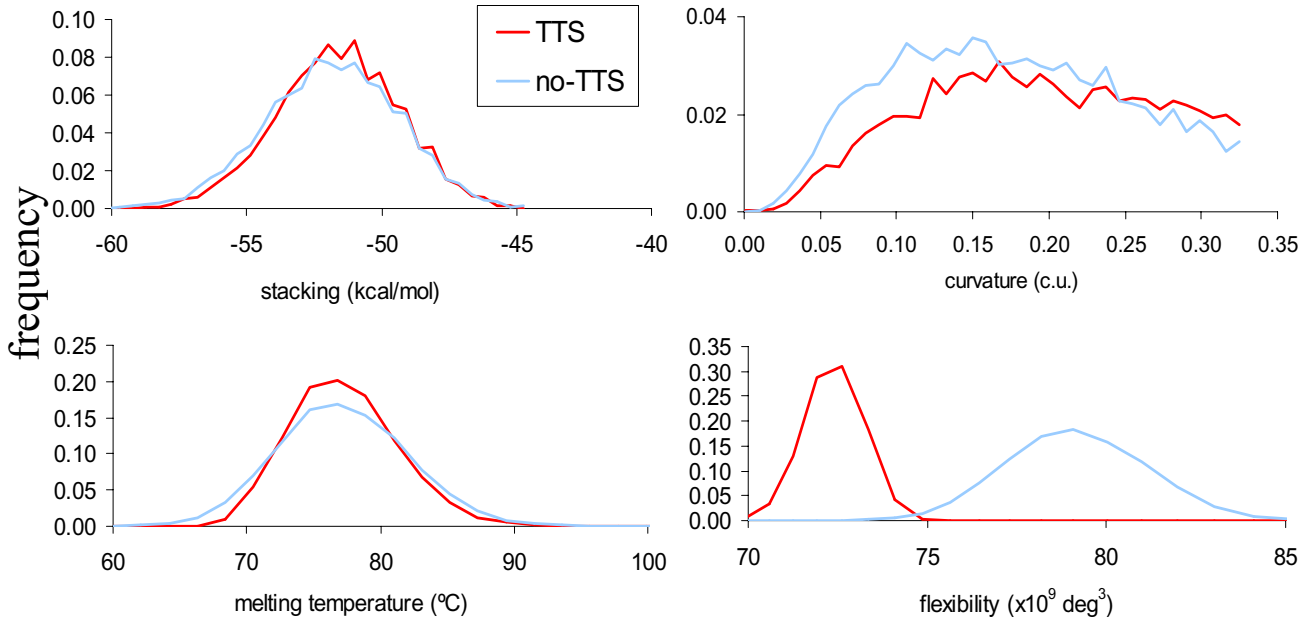

Figure 8
Distribution of selected physical properties of TTSs and random DNA sequences [see *Methods*] taken from human DNA. Analyses were performed on 25-nt fragments; similar results were obtained with shorter segments.

TTSs in transcription factor binding sites

The 0–200 nt region (specially the 0–100 segment) is expected to be largely enriched in transcription factor binding sites. We located all the putative transcription factor binding sites in these region of the human genome by mapping the last public version of the TRANSFAC database [34] to the *upstream5000* file in the UCSC-Genome Database [28] using the TFBS Perl module [35]. We then computed the average percentage of nucleotides in a TTS with a length of 10 or greater as moving apart the centre of the transcription factor recognition sequence.

Physical descriptors of DNA

DNA curvature calculation can be done using the data and the algorithm developed by Bolshoy and co-workers [40]. This algorithm calculates the three-dimensional path of a DNA molecule and estimates the curvature of the axis path. The scale is in arbitrary curvature units (c.u.), ranging from 0 (e.g. no curvature) to 1.0, which is the curvature of DNA when wrapped around the histone core of nucleosome.

To predict the stability of the sequence, we used the base step data from Santalucia *et al*. and the formula described in their study [41]. DNA stacking energies are predicted using the accurate interaction energies published for Sponer *et al*. [42,43] for nucleic acid base pairs.

The flexibility of a track was measured by the configurational volume [see eq. 1 in reference 44] in function of the Twist-twist, Shift-shift and Roll-roll force constants determined from MD simulations by Lankas *et al*., [45].

$$V = \sqrt{\frac{(kT)^3}{Ktwist \times Ktilt \times Kroll}} \qquad \text{eq. 1}$$

where k is Boltzman constant, T is the temperature (taken as 298 K) and *Ktwist*, *Ktilt* and *Kroll* are harmonic force constants expressed in kcal/mol • deg^2

Described methods are implemented in a Perl script library, witch is available upon request.

Authors' contributions

JRG performed most of the analysis presented in this paper. JMV performed the GO-analysis, with the help and advice of JD. MO performed the analysis of data, directed the study and wrote the paper.

Additional material

Additional File 1

Frequencies of nucleotides forming part of TTSs for different lengths in the human genome and for random models. Genome, promoter and promoter shuffled data is the same as in Figure 2. Genome and promoter random are computed using a numerical method [see Methods*] that maintains the trimer (or dimer) population.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-63-S1.pdf]

Additional File 2

A) Percentage of genes with an annotated CpG island in promoter region for a given GO term B) Differential GO-analysis for Genes with TTS of length 20 at the 100 upstream region of promoter. Left panel for the bulk of genes (identical to profiles (100-b) in Figure 5, middle for genes with CpG island, and right for genes without CpG genes.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-63-S2.pdf]

Additional File 3

Percentage of human-mouse identity of 15-nt fragments in several promoter regions for TTS and non-TTS segments of the same size in regulatory region. Alignments used here (difference with Figure 7) were taken from UCSC multiz8way data file. TFBS line show data for predicted transcription factor binding sites [see Methods*] in every region. 100 non-overlapping random sampling of non-TTS set is computed to calculate error bars.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-7-63-S3.pdf]

Acknowledgements

We thank Drs. Roderic Guigó and Xavier de la Cruz for their helpful discussion. This work was supported by the *Fundació La Caixa*, *Fundación BBVA* and the Spanish Ministry of Science (BIO2003-06848).

References

1. Fesenfeld G, Davis DR, Rich A: **Formation of a three-stranded polynucleotide molecule.** *J Am Chem Soc* 1957, **79:**2023-2024.
2. Soyfer VN, Potaman VN: *Triple-Helical Nucleic Acids* Springer-Verlag: New York; 1996.
3. Scaria PV, Shafer RH: **Calorimetric analysis of triple helices targeted to the $d(G_3A_4G_3)\cdot d(C_3T_4C_3)$ duplex.** *Biochemistry* 1996, **35:**10985-10994.
4. Robles J, Grandas A, Pedroso E, Luque FJ, Eritja R, Orozco M: **Nucleic acid triple helices: stability effects of nucleobase modifications.** *Curr Org Chem* 2002, **6:**1333-1368.
5. Chandler SP, Fox KR: **Specificity of antiparallel DNA triple helix formation.** *Biochemistry* 1996, **35:**15038-15048.
6. Shields GA, Laughton CA, Orozco M: **Molecular dynamics simulations of the d(T•A•T) triple helix.** *J Am Chem Soc* 1997, **119:**7463.
7. Jiménez E, Vaquero A, Espinás ML, Soliva R, Orozco M, Bernués J, Azorin F: **The GAGA factor of Drosophila binds triple-stranded DNA.** *J Biol Chem* 1998, **273:**24640.
8. Giovannangeli C, Hélène C: **Triplex technology takes off.** *Nature Biotechnology* 2000, **18:**1245.
9. Knauert MP, Glazer PM: **Triplex forming oligonucleotides: sequence-specific tools for gene targeting.** *Human MolecularGenetics* 2001, **10:**2243.
10. van Dongen MJP, Doreleijers JF, van der Marel GA, van Boom JH, Hilbers CW, Wijmenga SS: **Structure and mechanism of formation of the H-y5 isomer of an intramolecular DNA triple helix.** *Nature Structural Biology* 1999, **6:**854.
11. Strobel SA, Dervan PB: **Triple helix-mediated single-site enzymatic cleavage of megabase genomic DNA.** *Methods Enzymol* 1992, **216:**309.
12. Zain R, Marchand C, Sun J, Nguyen CH, Bisagni E, Garestier T, Hélène C: **Design of a triple-helix-specific cleaving reagent.** *Chem Biol* 1999, **6:**771.
13. Luo Z, Macris MA, Faruqi AF, Glazer PM: **High-frequency intrachromosomal gene conversion induced by triplex-forming oligonucleotides microinjected into mouse cells.** *Proc Natl Acad Sci USA* 2000, **97:**9003.
14. Havre PA, Gunther EJ, Gasparro FP, Glazer PM: **Targeted mutagenesis of DNA using triple helix-forming oligonucleotides linked to psoralen.** *Proc Natl Acad Sci USA* 1993, **90:**7879.
15. Majumdar A, Khorlin A, Dyatkina N, Lin FL, Powell J, Liu J, Fei Z, Khripine Y, Watanabe KA, George J, Glazer PM, Seidman MM: **Targeted gene knockout mediated by triple helix forming oligonucleotides.** *Nat Genet* 1998, **20:**212.
16. Barre FX, Ait-Si-Ali S, Giovannangeli C, Luis R, Robin P, Pritchard LL, Hélène C, Harel-Bellan : **Unambiguous demonstration of triple-helix-directed gene modification.** *Proc Natl Acad Sci USA* 2000, **97:**3084.
17. Wang G, Seidman MM, Glazer PM: **Mutagenesis in mammalian cells induced by triple helix formation and transcription-coupled repair.** *Science* 1996, **271:**802.
18. Vasquez KM, Narayanan L, Glazer PM: **Specific mutations induced by triplex-forming oligonucleotides in mice.** *Science* 2000, **290:**530.
19. Duval-Valentin G, Thuong NT, Hélène C: **Specific inhibition of transcription by triple helix-forming oligonucleotides.** *Proc Natl Acad Sci USA* 1992, **89:**504.
20. Cooney M, Czernuszewicz G, Postel EH, Flint SJ, Hogan ME: **Site-specific oligonucleotide binding represses transcription of the human c-myc gene in vitro.** *Science* 1988, **241:**456.
21. Grigoriev M, Praseuth D, Robin P, Hemar A, Saison-Behmoaras T: **A triple helix-forming oligonucleotide-intercalator conjugate acts as a transcriptional repressor via inhibition of NF kappa B binding to interleukin-2 receptor alpha-regulatory sequence.** *J Biol Chem* 1992, **267:**3389.
22. Joseph J, Kandala JC, Veerapanane D, Weber KT, Guntaka RV: **Antiparallel polypurine phosphorothioate oligonucleotides form stable triplexes with the rat alpha1(I), collagen gene promoter and inhibit transcription in cultured rat fibroblasts.** *Nucleic Acids Res* 1997, **25:**2182.
23. Postel EH, Flint SJ, Kessler DJ, Hogan ME: **Evidence that a triplex-forming oligodeoxyribonucleotide binds to the c-myc promoter in HeLa cells, thereby reducing c-myc mRNA levels.** *Proc Natl Acad Sci USA* 1991, **88:**8227.
24. Young SL, Krawczyk SH, Matteucci MD, Toole JJ: **Triple helix formation inhibits transcription elongation in vitro.** *Proc Natl Acad Sci USA* 1991, **88:**10023.
25. Faria M, Wood CD, Perrouault L, Nelson JS, Winter A, White MR, Hélène C: **Targeted inhibition of transcription elongation in cells mediated by triplex-forming oligonucleotides.** *Proc Natl Acad Sci USA* 2000, **97:**3862.
26. Goñi JR, de la Cruz X, Orozco M: **Triplex forming oligonucletide target sequences in the human genome.** *Nucleic Acids Res* 2004, **32:**354-360.
27. Anderson JD, Widom J: **Poly(dA-dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites.** *Mol Cell Biol* 2001, **11:**3830-3839.
28. Karolchick D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Kent WJ: **The UCSC Genome Browser Database.** *Nucleic Acid Res* 2003, **31:**51.
29. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
30. Rice P, Longden I, Bleasby A: **"EMBOSS: The European Molecular Biology Open Software Suite".** *Trends in Genetics* 2000, **16(6):**276-277.

31. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, Weng Z: **Detection of functional DNA motifs via statistical over-representation.** *Nucleic Acids Res* **32(4):**1372-81. 2004 Feb 26
32. **The NCBI Database** [http://www.ncbi.nlm.nih.gov/]
33. Myers G: **A Fast Bit-Vector Algorithm for Aproximate String Matching Based on Dynamics Programming.** *JACM* 1999, **46:**395-415.
34. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Prüβ M, Reuter I, Schacherer F: **TRANSFAC: an integrated system for gene expression regulation.** *Nucleic Acids Res* 2000, **28:**316-319.
35. Lenhard B, Wasserman WW: **TFBS: Computational framework for transcription factor binding site analysis.** *Bioinformatics* 2002, **18:**1135-1136.
36. Al-Shahrour F, Díaz-Uriarte R, Dopazo J: **FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes.** *Bioinformatics* 2004, **20:**578-580.
37. Al-Shahrour F, Minguez P, Vaquerizas J, Conde L, Dopazo J: **Babelomics: a suite of web-tools for functional annotation and analysis of group of genes in high-throughput experiments.** *Nucleic Acids Research* 2005 in press.
38. Gene Ontology Consortium: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res* 2004:D258-61.
39. Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple testing under dependency.** *Annals of Statistics* 2001, **29:**1165-1188.
40. Shpigelman ES, Trifonov EN, Bolshoy A: **CURVATURE: software for the analysis of curved DNA.** *Comput Appl Biosci* 1993, **9:**435.
41. Santalucia J Jr: **A unified view of polymer, dumbbell, and oligonucleotide DNA neares-neighbor thermodynamics.** *Proc Natl Acad SciUSA* 1998, **95:**1460.
42. Sponer J, Gabb HA, Leszczynski J, Hobza P: **Base-base and deoxyribose-base stacking interactions in B-DNA and Z-DNA: a quantum-chemical study.** *Biophys J* 1997, **73:**76-87.
43. Sponer J, Jurecka P, Hobza P: **Accurate interaction energies of hydrogen-bonded nucleic acid base pairs.** *J Am Chem Soc* 2004, **126:**10142-51.
44. Pérez A, Blas JR, Rueda M, López-Bes JM, de la Cruz X, Orozco M: **Exploring the essential dynamics of DNA.** *J Chem Theor Comput* 2005, **1:**790-800.
45. Filip Lankas , Jirí Sponer , Jörg Langowski , Thomas Cheatham E III: **DNA Basepair Step Deformability Inferred from Molecular Dynamics Simulations.** *Biophys J* 2003, **85:**2872.

Method

Determining promoter location based on DNA structure first-principles calculations

J Ramon Goñi[*†‡], Alberto Pérez[*†], David Torrents[§¶] and Modesto Orozco[*§¥]

Addresses: [*]Institute for Research in Biomedicine, Parc Científic de Barcelona, Josep Samitier, Barcelona 08028, Spain. [†]Departament de Bioquímica i Biología Molecular, Facultat de Biología, Avgda Diagonal, Barcelona 08028, Spain. [‡]Grup de recerca en Bioinformàtica i Estadística Mèdica, Departament de Biologia de Sistemes, Universitat de Vic. Laura, 13 08500 VIC, Spain. [§]Computational Biology Program, Barcelona Supercomputer Center, Jordi Girona, Edifici Torre Girona, Barcelona 08028, Spain. [¶]Institut Català per la Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys, 23. Barcelona 08010, Spain. [¥]Instituto Nacional de Bioinformática, Structural Bioinformatics Unit, Parc Cientific de Barcelona, Josep Samitier, Barcelona 08028, USA.

Correspondence: Modesto Orozco. Email: modesto@mmb.pcb.ub.es

Published: 11 December 2007

Genome ***Biology*** 2007, **8:**R263 (doi:10.1186/gb-2007-8-12-r263)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/12/R263

Received: 12 September 2007
Revised: 24 November 2007
Accepted: 11 December 2007

Abstract

A new method for the prediction of promoter regions based on atomic molecular dynamics simulations of small oligonucleotides has been developed. The method works independently of gene structure conservation and orthology and of the presence of detectable sequence features. Results obtained with our method confirm the existence of a hidden physical code that modulates genome expression.

Background

Sequencing projects have revealed the primary structure of the genomes of many eukaryotes, including that of human as well as other mammals. Unfortunately, limited experimental data exist on the detailed mechanisms controlling gene expression; this dearth of data has largely arisen from the difficulties found in the identification of regulatory regions. Traditionally, the immediate upstream region (200-500 bps) of a transcribed sequence is considered the proximal promoter area, where the binding of multiple transcription factor proteins triggers expression [1]. Other regulatory signals are found in distal regions (enhancers) that, despite being very far away in terms of sequence base pairs, can interact with the pre-initiation complex through the chromatin quaternary structure [1].

From a naïve perspective, the identification of promoter regions might be considered a trivial task, since they should be located immediately upstream (5') of the annotated transcribed regions. Unfortunately, the real situation is much more complex: on the one hand, 5' untranslated regions (UTRs) are very poorly described, and on the other, one gene might have several transcription start sties (TSSs) controlled by one or more proximal promoter regions (sometimes overlapping) scattered along gene loci, including introns, exons and 3' UTRs [2-6]. As a consequence, inspection of gene structure alone does not guarantee that the promoters will be located, and then, other signals need to be used to do this. Unfortunately such signals are very unspecific. Thus, transcription factor proteins are promiscuous and, depending on the genomic environment and the presence of alternative binding proteins, a given sequence can be recognized or ignored by the target protein. More general sequence signals also give noisy, unspecific signals. For example, the TATA box [7], which was originally believed to be associated with nearly all promoters, has been found to be present in only a small proportion of them [2,4]. A more powerful promoter signal stems from the presence of CpG islands [8-19], but even when

present their signal is rather diffuse and unspecific. In summary, promoter detection is one of the greatest experimental and computational challenges in the post-genomic era.

Current methods for promoter location are based on two approaches: the use of gene structure and conservation; and the existence of sequence profiles that might signal promoter region. In the first case, statistical algorithms are used to find signals of genes that locate the 5'-end and conserved regions upstream [20]. For the second case, many sequence/compositional rules haven been used. Thus, several algorithms have been developed to detect signals like the TATA box, CpG islands or regions with large populations of transcription factor binding sites (TFBSs) [1,12,13,16,21-28]. Compositional rules (from trimer to n-mer) have also been considered to enrich the differential signal at promoters [1,12,13,21-28]. Finally, some methods have used predicted gene structure [1,12,21,22,27-29] and its conservation across species [1,28,29] to help their sequence-trained models to locate promoters. However, despite recent progress, the performance of all these methods is not great, especially when used to predict promoters that are not part of canonical 5' upstream regions [5,11,15,23].

Clearly, diffuse factors other than the specific hydrogen-bond interactions between nucleotides and binding proteins modulate the recognition of target DNA fragments in promoter regions. As first suggested by Pedersen *et al.* [30], one of these additional factors can be the physical properties of DNA, which control the modulation of chromatin structure, the transmission of information from enhancers or proximal promoters, and the formation of protein aggregates in the pre-initiation complex. Thus, Pedersen and others have shown how some descriptors that are believed to be related to physical characteristics of DNA (such as DNase I susceptibility, A-phylicity, nucleosome preference, DNA stability, and so on, up to 15 strongly correlated descriptors [31]) can help to locate promoters in prokaryotes and, perhaps, in eukaryotes [14,30,32-35]. Recent versions of progams like *mcpromoter* [33] or *fprom* [1] have incorporated these parameters into their predictive algorithms [1,5,33].

In this paper, for the first time, we explore the possibility of using a well-defined physically based description of DNA deformability [36] derived from atomic simulations to determine promoter location. Parameters describing the stiffness of DNA were rigorously derived from long atomistic molecular dynamics (MD) simulations in water using a recently developed force-field fitted to high level *ab initio* quantum mechanical calculations [37]. Using exclusively these simple parameters, whose interpretation is clear and unambiguous, we developed an extremely simple predictive algorithm which performs remarkably well in predicting human promoters, even those located in unexpected genomic positions.

Results and discussion

Derivation of stiffness parameters of DNA from molecular dynamics simulations

The use of a recently developed force-field [37] allowed us to perform long MD simulations (50 ns) of different DNA duplexes from which parameters describing dinucleotide flexibility can be obtained. Trajectories are stable with the DNA maintaining a B-type conformation with standard hydrogen bond pairings (Figures S1 and S2 in Additional data file 1), no backbone deformations [37,38], and normal distributions on helical parameters (Figures S3 and S4 in Additional data file 1) centered on expected values.

In contrast to assumptions in ideal rod models, DNA deformability is largely dependent on sequence. For example, it is possible to unwind (with the same energy cost) a d(CG) step twice than a d(AC) one (see Table 1). Our analysis shows also

Table 1

Stiffness constants associated to helical deformations

Step	Twist	Tilt	Roll	Shift	Slide	Rise
AA	0.026	0.038	0.020	1.69	2.26	7.65
AC	0.036	0.038	0.023	1.32	3.03	8.93
AG	0.031	0.037	0.019	1.46	2.03	7.08
AT	0.033	0.036	0.022	1.03	3.83	9.07
CA	0.016	0.025	0.017	1.07	1.78	6.38
CC	0.026	0.042	0.019	1.43	1.65	8.04
CG	0.014	0.026	0.016	1.08	2.00	6.23
GA	0.025	0.038	0.020	1.32	1.93	8.56
GC	0.025	0.036	0.026	1.20	2.61	9.53
TA	0.017	0.018	0.016	0.72	1.20	6.23

Constants related to rotational parameters are in kcal/mol degree^2, while those related to translations are in kcal/mol Å^2.

that some steps are universally flexible (like d(TA)), while others are, in general, rigid (like d(AC)). However, the concept of 'stiffness' associated with a step is often meaningless, since depending on the nature of the helical deformation, the relative rigidity of two steps can change (Table 1). In summary, flexibility appears as a subtle-sequence dependent process that is quite difficult to represent without the help of powerful techniques like MD simulations.

Differential physical properties of human promoters

From the analysis of helical stiffness along the human genome (see parameters in Table 1 and Materials and methods), we detected regions with distinctive structural properties that show a strong correlation with annotated TSSs (located using the 5' end of the human Havana gene collection [39] in the Encode region [40]). In particular, this signal was significantly stronger in regions located from -250 bp to +900 bp of the TSSs (that is, covering the core and proximal promoter regions; Figure 1), which agrees with the particular structural needs attributed to the correct function of regulatory regions. Interestingly, the differential signal found at the genome-scale does not appear to depend exclusively on the presence of CpG islands since the same signature is also present (even with less intensity) in promoters with standard CpG content (Figure 1c,d). Compared to those regions that are located far from annotated TSSs, the structural pattern measured for regulatory regions is quite complex: high flexibility near TSSs is required for some parameters, while rigidity is needed for others (Figure 1). Thus, our results suggest that the pattern of flexibility needed in promoter regions is quite unique, and general concepts like 'curvature propensity' or 'general flexibility' are too simplistic to capture the real average physical properties of promoter regions. We can speculate that the need for proper placement of nucleosomes, combined with the specific structural requirements of multiprotein complexes, favor the presence of sequences with unique deformation properties in the promoter region (especially in the core and proximal regions), which can be measured computationally.

Using structural parameters for promoter prediction: ProStar

Taking advantage of the specific pattern of flexibility of promoter regions described above, we developed a new predictive algorithm called ProStar (for Promoter Structural Parameters; see Materials and methods), which uses only descriptors derived from physical first-principle type calculations (Table 1) to locate promoter regions (including strand orientation). Our method is conceptually and computationally simpler than any other general promoter prediction algorithm as it does not require any additional information, such as conservation of gene structure across species, presence of CpG islands, TATA-boxes, Inr elements or any other sequence specific signals. Due to its simplicity, ProStar can, in principle, be applied even in cases where promoters are located in unusual genomic positions.

In order to evaluate the performance of our methodology in the context of other promoter predicting approaches (see Materials and methods and Table S1 in Additional data file 2), we compared our results with those derived from other reported promoter predictors, following the Egasp workshop procedures [5,41] and using the annotation of the Havana team [39] for the Encode regions [40] as the reference set. In order to cover the whole spectrum of prediction methodologies, we selected a few representative procedures mainly based on the conservation of gene structure (*fprom* [1], *firstef* [13], *dpf* [12] and *nscan* [29]), the identification of CpG islands (*eponine* [22], *cpgprod* [16] and *dgsf* [21]), compositional sequence biases (*mcpromoter* [26,33]) and other criteria (*nnpp* [24] and *promoter2.0* [25]). The results of these comparisons show that despite its simplicity, ProStar performed better than most of the other methods and was similar to two algorithms that use gene structure for prediction (*fpom* and *firstef*), and only *nscan*, which is based also on multi-species homology, provided more accurate results for the reference set of genes (Figure 2, Table 2 and Figure S5 in Additional data file 1). Global analysis of performance using Bajic's metrics [42] (see Materials and methods) showed that the predictive power of our method is only improved by *nscan* (Table 2 and Table S2 in Additional data file 2). Furthermore, when the calculations used to derive the results shown in Figure 2 are repeated using a more restrictive tolerance test (window size D = 250; see Materials and methods), the superiority of ProSart with respect to most of the other methods was maintained (Figure S6 in Additional data file 1) in most regions of a 'proportion of correct predictions (PPV)/sensitivity (SENS)' map, demonstrating the robustness of our method. Finally, it is worth to comment the good performance of ProStar, that only uses simple dinucleotide parameters, compared to complex methods based on n-mer compositional rules (see Materials and methods). Clearly, the richness of the six-dimensional descriptors obtained for each dinucleotide by the MD simulation explains the success of our simple approach.

Interestingly, when the analysis is performed for a subset of TSSs of non-coding genes (Figure 2, Table 2 and Figure S6 in Additional data file 1) the performance of all the methods decreases, but ProStar seems more robust than the others. In fact, the analysis of these data shows that, for this subset of genes, ProStar performs better than any method that uses sequence compositional bias, location of known TFBSs, or the presence of TATA-box signals or CpG islands and similar or better than those relying on the presence of orthologs as shown in Bajic's metrics (Table 2).

Testing ProStar against non-trivially identified promoters

Our method works better when predicting promoters associated with CpG islands, but the decrease in performance for promoters associated with non-CpG islands is similar to that of other methods, including those that are based on the main-

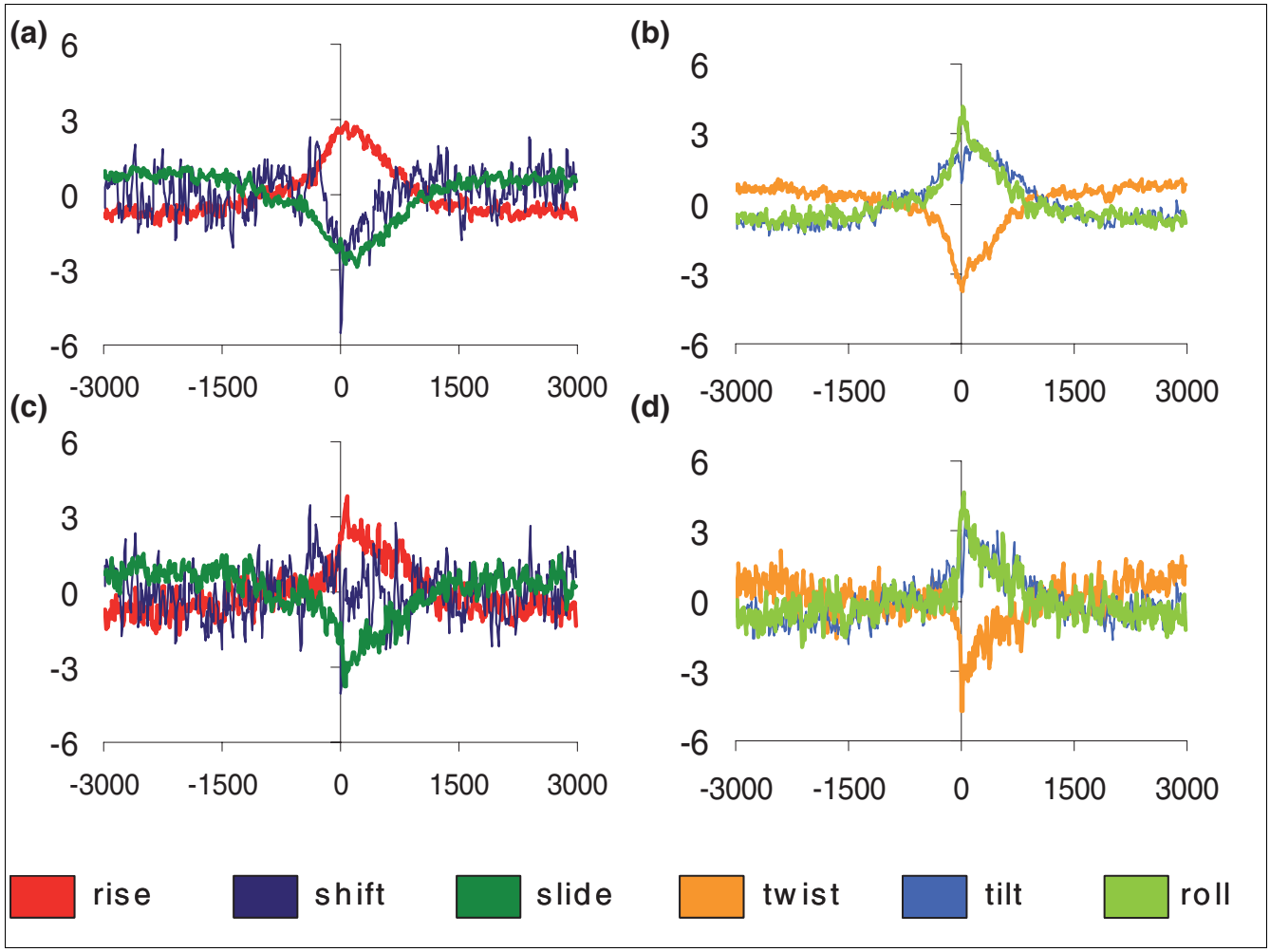

Figure 1
Measurement of the six 'average' helical force-constants. **(a,c)** Rise, shift, and slide; **(b,d)** twist, tilt, and roll. Results are shown for the complete training set of promoter regions (a,b) (see Materials and methods) and for the subset with no CpG island (c,d). Sequences are aligned at point +1 by its annotated TSS. All values are centered at zero (the background values).

tenance of the gene structure (Figure S7a in Additional data file 1). If a conservative definition of a non-CpG associated promoter is used (no CpG island detectable at less than 5 Kb from the promoter), the performance of ProStar decreases, but is still better than that of most methods (Figure S7b in Additional data file 1), although even in this case the method is not competitive with algorithms based on gene structure conservation. In any case the performance of ProStar for genes not associated with CpG islands is quite reasonable, confirming that the need for specific elastic properties at promoter regions is a general requirement and not restricted to the presence of CpG islands or diffuse TSSs. It is also worth noting that ProStar performs better than methods specifically tuned to capture promoters associated with CpG islands when the analysis is restricted to Havana annotated genes with CpG islands (data not shown). Finally, the performance of ProStar does not decay for genes containing a TATA box (Figure S8 in Additional data file 1), which are the easiest to detect from simple sequence signals.

Once we tested the performance of ProStar to reproduce promoters annotated by the Havana group, we explored the ability of the method to locate promoters reported in massive Cage experiments [4], where promoters were often found in unexpected locations. To increase the challenge, we analyzed only Cage-detected promoters falling inside transcribed regions (including exons and 3' UTR regions) of annotated Havana genes that are not regulated by a CpG island. Our results demonstrate that despite the method not being trained with this type of promoter, it performed quite well (Figure 2, Table 2, Figures S6 and S9 in Additional data file 1), in fact improving the results obtained by other available methods (Table 2).

Table 2

Global ASM performance index obtained by considering Bajic's muti-metric analysis for different sets of genes

	CDS_gene	no_CDS_gene	noCpG	no_CpG_CAGE
ProStar	2.78	**2.00**	6.56	**2.56**
cpgprod	8.22	7.89	7.22	7.11
dgsf	9.56	9.11	9.11	7.00
dpf	6.78	7.00	4.89	5.67
eponine	5.56	6.11	8.33	3.78
firstef	4.00	4.00	5.56	3.78
fprom	3.56	3.22	2.78	9.78
mcpromoter	5.56	5.33	4.89	4.89
nnpp	10.44	10.33	9.44	8.89
nscan	**1.56**	2.89	**1.22**	6.89
promoter2.0	10.67	10.56	8.89	9.33
proscan	9.33	9.56	9.11	8.33

Global ASM performance index obtained following Bajic's muti-metric analysis (see Materials and methods) for different sets of genes: the 2,641 TSSs from the Havana set (column CDS_gene), the 1,764 TSSs of non-coding genes from the Havana set (column no_CDS_gene), the 1,751 TSSs of the Havana set that do not overlap any CpG island (column noCpG), and the collection of 1,086 Cage TSSs not associated with CpG islands (no_CpG_CAGE). In each case the method providing the best results is shown in bold. Note that ProStar is the best in the two most difficult categories and the second best over the entire set of genes.

ProStar calculations were repeated throughout the entire human genome using TSS positions according to RefSeq genes. The results are summarized in Figure S10 in Additional data file 1 and confirm the quality of our predictions at the genome level. Please note that some caution is needed in the interpretation of these results since the apparent better performance of our method at the genome level compared with that obtained using Encode regions can be simply due to the noise in the first dataset.

The final extreme challenge for ProStar was to find promoters that are not detectable by methods based on sequence conservation along orthologs or on the maintenance of gene structure. For this purpose, we selected a subset of 1,203 annotated promoters of non-coding genes that are found as false negative by *nscan*, *fprom* and *firstef*. We should clarify that this comparison will give no information on ProStar with respect to 'state of the art' methods based on conservation of gene structure and orthology, but does give some indication of the ability of other methods (including ProStar) to capture promoters located in anomalous positions. The results shown in Figure 3 demonstrate that ProStar can recover a significant fraction of these promoters with a signal to noise ratio superior to all methods based on the differential genomic content of promoters and on the use of powerful discriminant algorithms. This suggests that ProStar is a powerful tool for promoter determination and that it could be a good alternative for the location of promoters of fast evolving genes or those appearing in anomalous positions that violate the traditional concept of gene structure.

Conclusion

Atomic MD simulations, based on physical potentials derived from quantum chemical calculations, yield helical stiffness parameters that reveal the complexity of the deformation pattern of DNA. The use of these intuitive parameters at the genomic level allowed us to define promoters as regions of unique deformation properties, particularly near TSSs. Taking advantage of this differential pattern, we trained a very simple method, based on Mahalanobis metrics, that is able to locate human promoters with remarkable accuracy. Our results are better than the ones of methods based on the use of large batteries of descriptors, such as sequence signals, empirical physical descriptors, and complex statistical predictors (neural networks, hidden Markov models, and so on). The overall performance of ProStar is similar and in some cases even better than that of methods based on the conservation of gene structure, methods that might not be so accurate in the location of promoters of fast evolving genes, or those located in unusual positions. Taken together, our work reveals that even in complex organisms like human, there is a hidden physical code that contributes to the modulation of gene expression.

Materials and methods

Molecular dynamics simulations

In order to have enough equilibrium samplings for all the ten unique steps of DNA, we performed MD simulations of four duplexes containing several replicas of every type of base step dimer (d(GG), d(GA), d(GC), d(GT), d(AA), d(AG), d(AT), d(TA), d(TG) and d(CG)): d(GCCTATAAACGCCTATAA), d(CTAGGTGGATGACTCATT), d(CACGGAACCGGTTCCGTG) and d(GGCGCGCACCACGCGCGG). All duplexes were

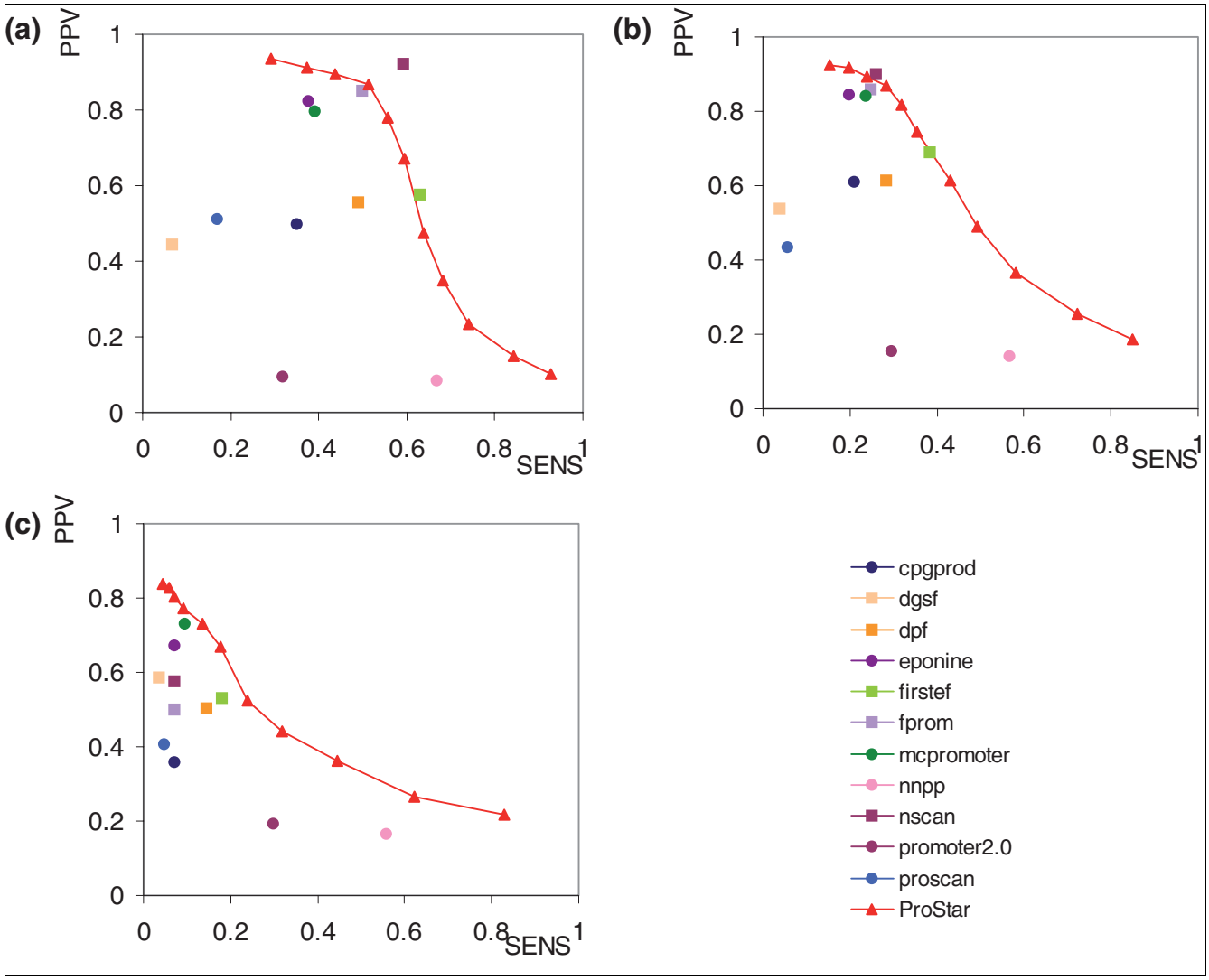

Figure 2
Results of performance comparison for the Encode region between ProStar and other programs (Table S1 in Additional data file 2) using a window size D equal to 1,000 (see Materials and methods). Results obtained compare the predictive power with **(a)** a subset of 885 Havana protein coding genes, **(b)** a set of 1,764 non-coding genes, and **(c)** a set of 1,086 annotated TSSs from a Cage data set that falls inside non-CpG island coding genes (see Materials and methods). Squares indicate methods based on gene prediction (exons, intronic signals, and so on), and other methods are represented with circles.

created in the standard B-type conformation, hydrated with around 10,600 water molecules, and neutralized by adding a suitable number of Na^+ ions. Neutral hydrated systems were then optimized, thermalized and pre-equilibrated using our standard protocol [43,44]. The structures obtained at the end of this procedure were then re-equilibrated for an additional 2 ns. The snapshots obtained at the end of this equilibration were used as starting points for 50 ns trajectories performed at constant temperature (298 K) and pressure (1 atm) using periodic boundary conditions and Ewald summations [45]. Simulations were carried out using SHAKE [46] on all bonds connecting hydrogens and 2 fts time steps for integration of Newton equations of motions. TIP3P [47] was used to represent water, while PARMBSC0 [37,48,49] was used to represent DNA.

Trajectories were manipulated to obtain the stiffness matrix (Ξ; equation 1) representing the deformability of a given step along rotations (twist, roll and tilt) and translations (rise, slide and shift) from equilibrium values. For this purpose we determined the oscillations of all these parameters, building a covariance matrix whose inversion led to the stiffness matrix (equation 1) [36,50-53], which is simplified for each dinucleotide step as a six-dimensional vector $\kappa = (k_{twist}, k_{roll}, k_{tilt}, k_{rise}, k_{shift}, k_{slide})$ by neglecting the out-of-diagonal terms in the stiffness matrix (equation 1). Note that each of these elements (k_i) is the force-constant associated with the distortion along a given helical coordinate:

$$\Xi = \left(k_B T\right)^{-1} \bullet C_h^{-1} = \begin{bmatrix} k_{twist} & k_{t-r} & k_{t-l} & k_{t-i} & k_{t-s} & k_{t-d} \\ k_{t-r} & k_{roll} & k_{r-l} & k_{r-i} & k_{r-s} & k_{r-d} \\ k_{t-l} & k_{r-l} & k_{tilt} & k_{l-i} & k_{l-s} & k_{l-d} \\ k_{t-i} & k_{r-i} & k_{l-i} & k_{rise} & k_{i-s} & k_{i-d} \\ k_{t-s} & k_{r-s} & k_{l-s} & k_{i-s} & k_{shift} & k_{s-d} \\ k_{t-d} & k_{r-d} & k_{l-d} & k_{i-d} & k_{s-d} & k_{slide} \end{bmatrix} \quad (1)$$

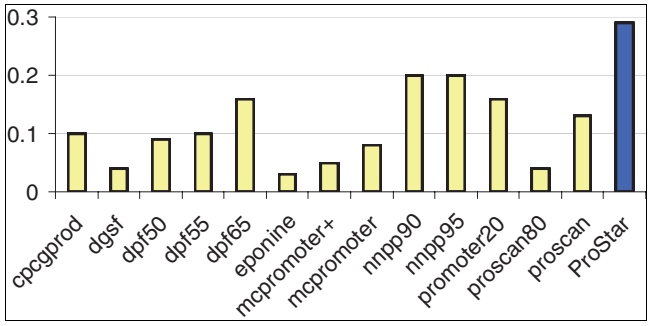

Figure 3
CC measurement (see Materials and methods) for the subset of Havana TSSs (1,203) of non-coding protein genes in the Encode region, unrecalled by *nscan*, *fprom* and *firstef*.

where k_B is Boltzman's constant, T is the absolute temperature and C_h is the covariance matrix in helicoidal space (for a given base step pair) obtained from the MD samplings.

Datasets

ProStar was trained using 5' ends of protein coding genes annotated by the Havana group [39] in the human Encode [40] region as a TSS set. According to Egasp workshop rules [5], the training procedure was restricted to 13 of the 44 Encode regions (see performance test section). TSS and strand recognition are trained and processed independently. ProStar requires a sequence with a minimum length of 500 nucleotides for TSS identification (see TSS prediction section). This size is extended to 1,800 nucleotides for strand prediction (see Strand prediction section).

Encode regions and annotated data and predictions were downloaded from the Egasp ftp directory [54]. We used *version00.3_20may* [55] of the Havana annotation and 'submitted_predictions' of the *egasp_submissions_20050503* directory [56] as predicted TSSs (Table S1 in Additional data file 2). The number of Havana TSSs that fall inside the Encode region is 2,641, but only 885 (34%) are coding genes. Coding genes are those with annotated start and stop codon signals; the others are taken as non-coding.

In addition to Egasp test sets, we analyzed the performance of our methodology using the selected sets of TSSs more difficult to predict (as TSSs on unexpected positions or TSSs belonging to genes with special particularities). These sets are a particular subset of 1,764 TSSs of Havana annotated non-coding genes (67% of Havana TSSs), 1,751 TSSs of coding and non-conding genes without upstream CpG islands (66% of the Havana set), 2,255 TTSs missing a TATA-box (85%), and the 1,086 unexpected TSSs positioned inside introns or exons of coding genes without CpG islands, as found in Cage predictions. CpG islands were mapped according to the UCSC database [57,58]. Since CpG islands are supposed to be the strongest promoter signals, this set represents an important challenge for our method. TATA-boxes were scanned in the proximal 50 nucleotide upstream region relative to the TSS, using the TATA position weight matrix [59] and the standard cut-off (-8.16). Cage predictions [60] were downloaded from Egasp [54] database. Those overlapping any Havana coding and non-coding genes (without a CpG island in the upstream region) were selected. Standard Egasp rules were used also for these challenging sets.

Training

We trained our method for promoter recognition with a collection of 500-nucleotide sequences that comprised intervals of 250 nucleotides upstream and downstream of the training TSS set. As negative set, we collected 500-nucleotide sequences from transcribed regions of Havana coding genes. We made sure that positive and negative sequences did not overlap. For the recognition of the strand, we trained our method with a collection of DNA sequences that comprised (for every TSS in the positive training set) the 1,800 nucleotide DNA sequence ranging from 900 bp upstream to 900 bp downstream of the same TSS. The reverse complementary sequences of the positive set were taken as a negative set.

Computation of DNA physical properties

Using our MD derived parameters (see Molecular dynamics simulations section and Table 1), we can describe any DNA sequence of size *n* as a six-dimensional deformation vector *v* = (*twist*, *tilt*, *roll*, *shift*, *slide*, *rise*). For a given deformation we sum the values associated with every dinuecleotide step in the sequence and divide the total by *n* - 1. For example, the *twist* deformation score for the sequence ACGC would be (0.036 [AC] + 0.014 [CG] + 0.025 [GC])/3 = 0.025. The six-dimensional vector of the same sequence would then be *v*(ACGT) = (0.025, 0.033, 0.022, 1.200, 2.547, 8.230).

Transcription start site prediction

We used Mahalanobis distance [61] to classify 500-nucleotide DNA sequences as belonging to the promoter class (k_x) or non-promoter class (k_y). Every class is defined by a specific dataset of sequences (see Training set section). Computing the physical properties of every sequence of the dataset, we conclude with a set of vectors for every class (*X* for class k_x and *Y* for k_y). The Mahalanobis distance D_M between the set of vectors *X* and *Y* is defined as:

$$D_M(X, Y) = (\mu_x - \mu_y)^t C^{-1}(\mu_x - \mu_y) \quad (2)$$

where μ_x and μ_y are the average vectors of the sets *X* and *Y* and C^{-1} is the covariance matrix of *XUY*. The decision function *g* of a specific 500-nucleotide DNA sequence with a descriptor vector *s* to a class k_i (with $i = <x, y>$) is defined as:

$$g(s,k_i) = w_{k_i}^t s + w_{k_i,0} \quad (3)$$

where $w_{ki} = C^{-1}\mu_i; w_{k_i,0} = -0.5\mu_i^t C^{-1}\mu_i$. When $g(s, k_x) > g(s, k_y)$ we should classify our sequence as a promoter. Even so, we can modulate the confidence of our decision according to a normalized score defined in equation 4. If the score is greater than a specific threshold (set to +1 by default), then the sequence is flagged as a promoter.

$$score(s) = \frac{g(s,k_X) - g(s,k_Y)}{g(\mu_X,k_X) - g(\mu_X,k_Y)} \quad (4)$$

Strand prediction

ProStar has been trained to recognize upstream/downstream signal asymmetry of predicted TSSs using a statistical discriminator based on Mahalanobis metrics (see last section) and on the differences in physical properties between the 0→-900 nucleotide and the 0→+900 nucleotide regions. The ProStar strand recognition module was trained using 1,800-nucleotide sequences with a TSS in the +900 position as the positive set. The reverse complement of the positive set sequences was used as the negative set.

Prediction clustering

As observed using experimental approaches [4], TSSs have a dominant position, but many closely related alternative sites may be found around them. In consequence, every TSS may produce multiple close predictions. To clarify the annotation, our algorithm allows the user to define a window size (set as 1,000 nucleotides by default) where all predictions will be unified in a single annotation. Accordingly, for a given window *W* of a specific strand *q*, we define *P*(*W*, *q*), the set of positions *p* falling inside *W* with $score(p, q) \geq c$ (where *c* is the user-defined minimal cutoff). Predicted dominant position *p'* of the window *W* is computed as:

$$p' = \frac{\sum_p p \cdot score(p,q)^2}{\sum_p score(p,q)^2} \quad (5)$$

Performance test

The training and performance of ProStar followed the protocol described [5] for the Egasp workshop [54,56]. Thus, protein coding genes annotated by the Havana group from 13 of the Encode regions were used for training, while the entire set was used in tests (tests performed using only regions that were not considered in the training give very close results; Table S2 in Additional data file 2). Also following the Egasp rules, true positives (TPs) are considered when the predicted TSS is in the same strand and at a maximum distance of D nucleotides from the annotated TSS (as in Egasp, D = 1,000 or D = 250 is used here). If the annotated TSS is missed using this criteria, we label the prediction as a false negative (FN). Every other prediction falling on the annotated part of the gene loci in the segment [+D+1, EndOfTheGene] counts as a false positive (FP). A true negative (TN) is the sum of positions falling on the gene loci segment [+D+1, EndOfTheGene] that do not overlap accepted true positive positions or any false positive prediction.

Sensitivity (SENS), proportion of correct predictions (PPV) and correlation coefficient (CC) are computed as:

$$SENS = \frac{TP}{TP+FN} \quad (7)$$

$$PPV = \frac{TP}{TP+FP} \quad (8)$$

$$CC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TF+FN)(TN+FP)(TN+FN)}} \quad (9)$$

In addition to the standard performance measures noted above, we also consider the average mismatch of predictions (AE) [5] and other extended metrics suggested by Bajic [42], including specificity (SPEC), Yule's association coefficient (Q), second prediction quality coefficient (K2), and generalized distances from ideal predictors (GDIP1, GDIP2, GDIP3). We also include in our analysis the averaged score measure (ASM), which combines many 'independent' descriptors to provide an overall relative measure of the quality of a predictive method with respect to others (Table S2 in Additional data file 2; Additional data file 3).

In addition to the methods checked in the Egasp experiment, we performed predictions using programs that were not considered in the Egasp experiment, but which are publicly available. In these cases we used the corresponding web-based tool or downloadable script with default parameters (Table S1 in Additional data file 2). When possible, we modified these default parameters in the input to obtain PPV/SENS curves (see Results and Figure S6 in Additional data file 1) instead of a single prediction. All methods were evaluated following the same thresholds for annotation of positive and negative predictions (see above).

Web server

ProStar is developed in C and compiled on a Linux machine. An unrestricted user-friendly version of the program is publicly available through our web server [62]. Strand prediction of recognized TSSs is an optional feature. Goodness of predictions may be tuned using a threshold (set to 1.0 by default) that may be increased to improve the proportion of correct predictions or decreased for sensitivity. Finally, the user may choice cluster size (see Prediction clustering section), which is

set to 1,000 by default. Clustering may be avoided by setting this size to small values (for example, 1).

Abbreviations

ASM, averaged score measure; CC, correlation coefficient; FN, false negative; FP, false positive; MD, molecular dynamics; PPV, proportion of correct predictions; SENS, sensitivity; SPEC, specificity; TFBS, transcription factor binding sites; TN, true negative; TP, true positive; TSS, transcription start sties; UTR, untranslated regions.

Authors' contributions

RG developed the predictive code and trained the method. AP performed the MD simulations and obtained the stiffness parameters. DT was involved in the design of the experiments and discussion of results and corrected the manuscript. MO conceived and developed the idea, designed and discussed experiments and wrote the manuscript.

Additional data files

The following additional data are available with the online version of this paper. Additional data file 1 provides supplementary figures showing plots of dinucleotide helical parameters and additional performance tests of ProStar. Additional data file 2 contains a list of promoter prediction methods described in this paper and a detailed evaluation of their performance. Additional data file 3 extends the description of the performance test and explains the averaged score measure (ASM)

Acknowledgements

This work has been supported by the Spanish Ministry of Education and Science (BIO2006-01602, BFU2004-01282 and BIO2006-15036) and the National Institute of Bioinformatics (Structural Bioinformatics Node). Calculations were performed at the MareNostrum supercomputer at the Barcelona Supercomputer Center.

References

1. Solovyev VV, Shahmuradov IA: **PromH: promoters identification using orthologous genomic sequences.** *Nucleic Acids Re* 2003, **31:**3540-3545.
2. Kim TH, Barrera LO, Qu C, Van Calcar S, Trinklein ND, Cooper SJ, Luna RM, Glass CK, Rosenfeld MG, Myers RM, Ren B: **Direct isolation and identification of promoters in the human genome.** *Genome Res* 2005, **15:**830-839.
3. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, *et al.*: **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309:**1559-1563.
4. Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, *et al.*: **Genome-wide analysis of mammalian promoter architecture and evolution.** *Nat Genet* 2006, **38:**626-635.
5. Bajic VB, Brent MR, Brown RH, Frankish A, Harrow J, Ohler U, Solovyev V, Tan SL: **Performance assessment of promoter prediction on ENCODE regions in the EGASP experiment.** *Genome Biol* 2006, **7(Suppl I):**S3-S13.
6. Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, *et al.*: **Antisense transcription in the mammalian transcriptome.** *Science* 2005, **309:**1564-1566.
7. Breatchnach R, Chambon P: **Organization and expression of eucaryotic split genes coding for proteins.** *Annu Rev Biochem* 1981, **50:**349-383.
8. Gardiner-Garden M, Frommer M: **CpG islands in vertebrate genomes.** *J Mol Biol* 1987, **196:**261-282.
9. Larsen F, Gundersen G, Lopez R, Prydz H: **CpG islands as gene markers in the human genome.** *Genomics* 1992, **13:**1095-1107.
10. Smale ST, Kadonaga JT: **The RNA polymerase II core promoter.** *Annu Rev Biochem* 2003, **72:**449-479.
11. Zhang MQ: **Computational prediction of eukaryotic protein-coding genes.** *Nat Rev Genet* 2002, **3:**698-709.
12. Bajic VB, Seah SH, Chong A, Krishnan SPT, Koh JLY, Brusic V: **Computer model for recognition of functional transcription start sites in RNA polymerase II promoters of vertebrates.** *J Mol Gaph Mod* 2003, **21:**323-332.
13. Davuluri RV, Grosse I, Zhang MQ: **Computational identification of promoters and first exons in the human genome.** *Nat Genet* 2001, **29:**412-417.
14. Pedersen AG, Baldi P, Chauvin Y, Brunak S: **The biology of eukaryotic promoter prediction - a review.** *Comput Chem* 1999, **23:**191-207.
15. Hannenhali S, Levy S: **Promoter prediction in the human genome.** *Bioinformatics* 2001, **17:**S90-96.
16. Ponger L, Mouchiroud D: **CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences.** *Bioinformatics* 2002, **18:**631-633.
17. Ioshikhes IP, Zhang MW: **Large-scale human promoter mapping using CpG islands.** *Nat Genet* 2000, **26:**61-63.
18. Antequera F, Bird A: **Number of CpG islands and genes in human and mouse.** *Proc Natl Acad Sci USA* 1995, **90:**11955-11959.
19. Suzuki Y, Tsunoda T, Sese J, Taira H, Mizushima-Sugano J, Hata H, Ota T, Isogai T, Tanaka T, Nakamura Y, *et al.*: **Identification and characterization of the potential promoter regions of 1031 kinds of human genes.** *Genome Res* 2001, **11:**677-684.
20. Gross SS, Brent MR: **Using multiple alignments to improve gene prediction.** *J Comput Biol* 2006, **13:**379-393.
21. Bajic VB, Seah SH: **Dragon Gene Start Finder identifies approximate locations of the 5' ends of genes.** *Nucleic Acids Res* 2003, **31:**3560-3563.
22. Down TA, Hubbard TJ: **Computational detection and location of transcription start sites in mammalian genomic DNA.** *Genome Res* 2002, **12:**458-461.
23. Reese MG: **Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome.** *Comput Chem* 2001, **26:**51-56.
24. Knudsen S: **Promoter2.0: for the recognition of PolII promoter sequences.** *Bioinformatics* 1999, **15:**356-361.
25. Prestridge DS: **Predicting Pol II promoter sequences using transcription factor binding sites.** *J Mol Biol* 1995, **249:**923-932.
26. Ohler U, Liao GC, Niemann H, Rubin GM: **Computational analysis of core promoters in the *Drosophila* genome.** *Genome Biol* 2002, **3:**RESEARCH0087.
27. Solovyev V, Salamov A: **The Gene-Finder computer tools for analysis of human and model organisms genome sequences.** *Proc Int Conf Intell Syst Mol Biol* 1997, **5:**294-302.
28. Korf I, Flicek P, Duan D, Brent MR: **Integrating genomic homology into gene structure prediction.** *Bioinformatics* 2001, **17:**S140-148.
29. Brown RH, Gross SS, Brent MR: **Begin at the beginning: predicting genes with 5' UTRs.** *Genome Res* 2005, **15:**742-747.
30. Pedersen AG, Baldi P, Chauvin Y, Brunak S: **DNA structure in human RNA polymerase II promoters.** *J Mol Biol* 1998, **281:**663-673.
31. Ponomarenko JV, Ponomarenko MP, Frolov AS, Vorobyev DG, Overton GC, Kolchanov NA: **Conformational and physicochemical DNA features specific for transcription factor binding sites.** *Bioinformatics* 1999, **15:**654-668.
32. Pedersen AG, Jensen LJ, Brunak S, Staefeldt HH, Ussery DW: **A DNA structural atlas for *Escherichia coli*.** *J Mol Biol* 2000, **299:**907-930.
33. Ohler U, Nierman H, Liao GC, Rubin GM: **Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition.** *Bioinformatics* 2001, **17:**S199-206.
34. Kanhere A, Bansal M: **Structural properties of promoters: similarities and differences between prokaryotes and**

eukaryotes. *Nucleic Acids Res* 2005, **33:**3165-3175.
35. Florquin K, Saeys Y, Degroeve S, Rouze P, Van de Peer Y: **Large-scale structural analysis of the core promoter in mammalian and plant genomes.** *Nucleic Acid Res* 2005, **33:**4235-4264.
36. Olson WK, Gorin AA, Lu X, Hock LM, Zhurkin VB: **DNA sequence-dependent deformability deduced from protein-DNA crystal complexes.** *Proc Natl Acad Sci USA* 1998, **95:**11163-11168.
37. Pérez A, Marchán I, Svozil D, Sponer J, Cheatham TE, Laughton CA, Orozco M: **Refinement of the AMBER force field for nucleic acids. Improving the description of α/γ conformers.** *Biophys J* 2007, **92:**3817-3829.
38. Varnai P, Zakrzewska K: **DNA and its counterions: a molecular dynamics study.** *Nucleic Acids Res* 2004, **32:**4269-4280.
39. **The HAVANA Team** [http://www.sanger.ac.uk/HGP/havana/]
40. The ENCODE Project Consortium: **The ENCODE (ENCylopedia Of DNA Elements) Project.** *Science* 2004, **306:**636-640.
41. Reese MG, Guigó R: **EGASP: Introduction.** *Genome Biol* 2006, **7(Suppl 1):**S1-3.
42. Bajic VB: **Comparing the success of different prediction software in sequence analysis: a review.** *Brief Bioinform* 2000, **1:**214-228.
43. Shields GC, Laughton CA, Orozco M: **Molecular dynamics simulations of the d(T·A·T) triple helix.** *J Am Chem Soc* 1997, **119:**7463-7469.
44. Orozco M, Pérez A, Noy A, Luque FJ: **Theoretical methods for the simulation of nucleic acids.** *Chem Soc Rev* 2003, **32:**350-364.
45. Darden T, York D, Pedersen LG: **Particle Mesh Ewald: AN N-log(N) method for Ewald sums in large systems.** *J Chem Phys* 1993, **98:**10089-10092.
46. Ryckaert JP, Ciccotti G, Berendsen HGC: **Numerical-integration of Cartesian equations of motion of a system with constraints - molecular-dynamics of N-alkanes.** *J Comp Phys* 1977, **23:**327-341.
47. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML: **Comparison of simple potential functions for simulating liquid water.** *J Chem Phys* 1983, **79:**926-935.
48. Cheatham TE III, Cieplak P, Kollman PA: **A modified version of the Cornell e*t al*. force field with improved sugar pucker phases and helical repeat.** *J Biomol Struct Dyn* 1999, **16:**845-862.
49. Cornell WD, Cieplak P, Baily CI, Gould IR, Merz KM Jr, Ferguson DC, Fox T, Caldwell JW, Kollman PA: **A second generation force field for the simulation of proteins, nucleic acids, and organic molecules.** *J Am Chem Soc* 1995, **117:**5179-5197.
50. Lankas F, Sponer J, Langowski J, Cheatham TE: **DNA basepair step deformability inferred from molecular dynamics simulations.** *Biophys J* 2003, **85:**2872-2883.
51. Noy A, Pérez A, Márquez A, Luque FJ, Orozco M: **Structure, recognition properties and flexibility of the DNARNA hybrid.** *J Am Chem Soc* 2005, **127:**4901-4920.
52. Noy A, Perez A, Lankas F, Luque FJ, Orozco M: **Relative flexibility of DNA and RNA: a molecular dynamics study.** *J Mol Biol* 2004, **343:**627-638.
53. Pérez A, Noy A, Lankas F, Luque FJ, Orozco M: **The relative flexibility of DNA and RNA: Database analysis.** *Nucleic Acids Res* 2004, **32:**6144-6151.
54. **EGASP Meeting** [http://genome.imim.es/gencode/workshop/meeting.html]
55. **EGASP HAVANA Gene Annotation** [ftp://genome.imim.es/pub/projects/gencode/data/havana-encode/]
56. **EGASP Predictions** [ftp://genome.imim.es/pub/projects/gencode/data/egasp05/]
57. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas , *et al.*: **The UCSC Genome Browser Database.** *Nucleic Acids Res* 2003, **31:**51-54.
58. **UCSC Genome Browser** [http://genome.ucsc.edu/]
59. Bucher P: **Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences.** *J Mol Biol* 1990, **212:**563-578.
60. **EGASP CAGE TSS** [ftp://genome.imim.es/pub/projects/gencode/data/TSS_to_share/CAGE_Ditags_TSS.gff]
61. Marques de Sa JP: *Pattern Recognition: Concepts, Methods and Applications* Berlin: Springer Verlag; 2001.
62. **ProStar Web Server** [http://mmb.pcb.ub.es/proStar/]

BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 15 2008, pages 1731–1732
doi:10.1093/bioinformatics/btn259

Structural bioinformatics

DNAlive: a tool for the physical analysis of DNA at the genomic scale

J. Ramon Goñi[1,2], Carlos Fenollosa[1,2,3], Alberto Pérez[1,2,3,4], David Torrents[1,2,5] and Modesto Orozco[1,2,3,4,*]

[1]Joint IRB-BSC Program on Computational Biology, Institute of Research in Biomedicine, Parc Científic de Barcelona, Josep Samitier 1-5, Barcelona 08028, [2]Barcelona Supercomputing Center, Jordi Girona 31, Barcelona 08034, [3]National Institute of Bioinformatics, Parc Científic de Barcelona, Josep Samitier 1-5, [4]Departament de Bioquímica, Facultat de Biología, Avgda Diagonal 647, Barcelona 08028 and [5]Institut Català per la Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Received on March 27, 2008; revised on May 16, 2008; accepted on June 4, 2008

Advance Access publication June 9, 2008

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: DNAlive is a tool for the analysis and graphical display of structural and physical characteristics of genomic DNA. The web server implements a wide repertoire of metrics to derive physical information from DNA sequences with a powerful interface to derive 3D information on large sequences of both naked and protein-bound DNAs. Furthermore, it implements a mesoscopic Metropolis code which allows the inexpensive study of the dynamic properties of chromatin fibers. In addition, our server also surveys other protein and genomic databases allowing the user to combine and explore the physical properties of selected DNA in the context of functional features annotated on those regions.

Availability: http://mmb.pcb.ub.es/DNAlive/ ; http://www.inab.org/

Contact: modesto@mmb.pcb.ub.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Massive genomic projects have revealed the sequence of nearly 50 eukaryotic genomes, including several mammals (among them, humans) and many more will become available in the coming years. So far, the annotation of these genomes has been nearly restricted to the identification and the one-dimensional location of functional features (mostly genes and their regulatory regions), without considering the structural parameters of their environment, which have been proven to be crucial for the functionality of DNA. Determining the structural properties of DNA and the combination of functional features is necessary to interpret and understand the functionality of genomes in a more complex, and therefore real, environment. The identification of these structural parameters allows scientists to consider different levels of accessibility of certain DNA regions to different proteins, such as transcription factors, polymerases and DNA methylases. For example, specific deformability or helical properties in a given region of DNA facilitate or impair the formation of nucleosomes hundreds of base pairs away, or can affect dimerization of two DNA-binding proteins which might be separated by thousands of bases in sequence. Different groups (Abeel *et al.*, 2008; Goñi *et al.*, 2007; Ohler *et al.*, 2001; Pedersen *et al.*, 2000; Singhal *et al.*, 2008) have demonstrated that regulatory regions in DNA display unusual physical properties, and in fact, two groups have recently proven independently (Abeel *et al.*, 2008; Goñi *et al.*, 2007) that eukaryotic promoters can be located with surprisingly good accuracy just analyzing simple physical descriptors of DNA, which confirms the existence of a hidden physical code that controls gene function. In summary, functional annotation needs to be complemented with physical data to understand the structure, dynamics and the general functionality of genomic DNA.

DNAlive has been developed to give a complete description of the physical properties of genomic DNA in a simple way, thus providing data that can be easily understood by non-structural experts. Among others, DNAlive allows the user to (i) determine potential correlations between genome annotations (such as transcription start sites, exons, splicing sites, …) and a battery of 29 physical descriptors of DNA (stability, helical descriptors, curvature, non-canonical B-DNA affinity, stiffness, …); (ii) find out the most stable 3D structure of long genome fragments (both naked DNA and DNA-protein complexes) using sequence-dependent average helical parameters, and, when available, experimental structural data on DNA-protein complexes; (iii) perform a dynamic analysis of chromatin fiber exploring the range of deformability sampled during trajectory and the possibility of the formation of transient protein–protein complexes and (iv) display structural parameters of DNA in the context of associated functional features obtained form several public databases. The tool is available as a web page and also as different webservices, which can be incorporated in user workflows (Supplementary Material).

2 IMPLEMENTATION

2.1 Entry data

The only mandatory input data for DNAlive is a DNA sequence in FASTA format or the genomic coordinates of a supported

*To whom correspondence should be addressed.

vertebrate genome. The program retrieves parameters from their internal databases (Supplementary Table 1) to determine physical profiles and to create a 3D structure of the naked DNA. Given a DNA sequence, the program determines potentially bound transcription factor binding sites (TFBS) by scanning the public TRANSFAC database (http://www.gene-regulation.com/) linked to PDB (http://www.rcsb.org/) and Uniprot databases (http://www.ebi.uniprot.org/). The selection of the complex of interest can be monitored externally by the user, who can force the generation of specific complexes (for example, nucleosomes, protein-multicomplexes, etc.).

2.2 Server workflow

Once a DNA sequence is entered (Fig. 1), the program computes the profile for the 29 physical properties available for the fiber (Supplementary Table 1). All properties are represented in a 2D plot using either the UCSC Genome Browser (http://genome.ucsc.edu) in combination with annotated genes whenever genomic coordinates for the genome are provided, or Gnuplot (Fig. 1 and Supplementary Fig. 1).

To combine the visualization of DNA physical properties with public annotations of the genome, coordinates of the input DNA sequence can be matched by running a search in our local Blat server (Kent, 2002). Although the user is able to annotate transcription factor PDB structures on specific positions of the DNA input sequence, we have implemented an automatic method to perform this step using the TFBS Perl library (Lenhard and Wasserman, 2002). The reconstruction of the average 3D structure of DNA is achieved using sequence-dependent base step parameters derived from accurate atomistic molecular dynamics (Pérez, 2007) and making use of a local adaptation of X3DNA (Lu and Olson, 2003) script (Fig. 1 and Supplementary Fig. 2). When structural information on protein–DNA complexes is available, modeled structures in the corresponding segment are substituted by the experimental geometries, and junctions are refined if required. The visualization of 3D structures is performed by integrating Jmol Java applets (http://www.jmol.org/) in the HTML page. All physical descriptors can be mapped into the 3D structure to favor the detection of potential correlations between conformation, functional annotations and physico-chemical properties (Fig. 1).

The server also includes unique tools for a rapid representation of chromatin dynamics, which, in extensive analysis performed in our laboratory on our database of more than 100 trajectories, showed a surprisingly high accuracy of the essential deformation pattern of DNA. The method uses a mesoscopic Metropolis Monte Carlo algorithm, where the geometry of each base pair is defined by three local rotations (roll, tilt and twist) and translations (slide, shift and rise), and the conformational energy is estimated from the deformation matrix using a harmonic model (Equation 1), where the index 'i' stands for one of the M base pair steps and the index 'j' stands for the six unique helical parameters (ξ) for each step. The equilibrium values for one helical parameter in a given base pair step type and (ξ_{ij}^0) and the associated deformation constant ($K_{i,j}$) were previously determined from molecular dynamics simulations (Pérez, 2007). Once a movement in helical coordinates is accepted by the Metropolis test, the corresponding Cartesian structure of the fiber is generated using an adaptation of X3DNA (Lu and Olson, 2003) for VIDEO visualization using JMOL Java applets in the HTML page (Supplementary Fig. 3). Basic manipulation and analysis of the trajectories and structure (rotations, translations, distance measurements,…) are allowed by the Jmol interface, which allows the determination of potential DNA-mediated protein-clusters.

$$E = \sum_{i=1}^{M}\sum_{j=1}^{6} K_{i,j}\left(\xi_{ij} - \xi_{ij}^{0}\right)^2 \quad (1)$$

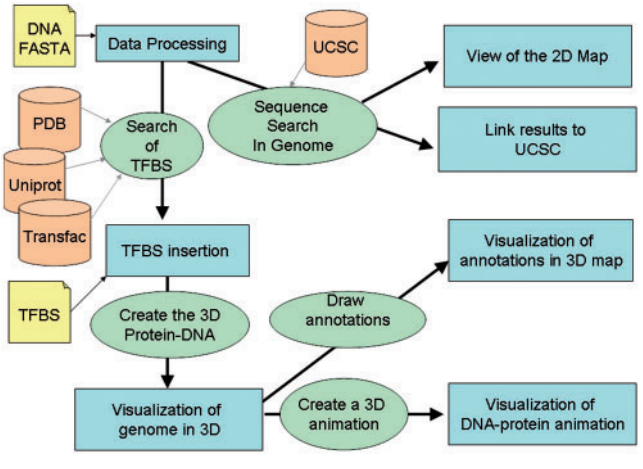

Fig. 1. DNAlive web server workflow diagram.

ACKNOWLEDGEMENTS

We thank the help of Agnes Noy, David Piedra, Henrique Proença and Joaquín Panadero as β-testers of the server.

Funding: This work has been supported by the Spanish Ministry of Education and Science (BIO2006-01602 and BIO2006-15036), the Spanish Ministry of Health (COMBIOMED network), the Fundación Marcelino Botín and the National Institute of Bioinformatics (Structural Bioinformatics Node).

Conflict of Interest: none declared.

REFERENCES

Abeel,T. *et al.* (2008) Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res.*, **18**, 310–323.

Goñi,J.R. *et al.* (2007) Determining promoter location based on DNA structure first-principles calculations. *Genome Biol.*, **8**, R263.

Kent,W.J. (2002) BLAT- the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.

Lu,X.J. and Olson,W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–5121.

Ohler,U. *et al.* (2001) Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics*, **17** (Suppl. 1), S199–S206.

Pedersen,A.G. *et al.* (2000) A DNA structural atlas for *Escherichia coli*. *J. Mol. Biol.*, **299**, 907–930.

Pérez,A. *et al.* (2007) Refinement of the AMBER force field for nucleic acids. Improving the description of α/γ conformers. *Biophys. J.*, **92**, 3817–3829.

Singhal,P. *et al.* (2008) Prokaryotic gene finding based on physicochemical characteristics of codons calculated from molecular dynamics simulations. *Biophys. J.* [EPub ahead of print; DOI:10.1529/biophysj.107.116392].