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

6 DISCUSSIÓ

En aquest capítol es discuteixen breument els resultats més rellevants obtinguts en el transcurs de la tesis. Es recomana adreçar-se als articles originals per una descripció més detallada dels estudis.

6.1 LES SEQÜÈNCIES FORMADORES DE TRÍPLEX EN EL GENOMA HUMÀ

Donat el gran nombre d'aplicacions les triple hèlices, han aparegut gran nombre d'estudis centrats en millorar la penetració cel·lular del TFO, l'accés al la seqüència diana (TTS) en el nucli i en la millora de l'estabilitat de la molècula mitjançant modificacions en els nucleòtids. Però malgrat aquest esforços, i tot i el que els tríplex estan molt ben caracteritzats i validats experimentalment, no existeix encara cap agent terapèutic basat en aquesta estratègia (Duca, Vekhoff, Oussedik, Halby, & Arimondo, 2008). Una de les possibles causes d'aquesta situació és el desconeixement de quin és el potencial real del genoma humà per formar tríplexs en presència de TFOs i com una potencial formació de tríplex pot alterar l'estructura de l'ADN, i per tant la seva funcionalitat, en regions claus per la regulació.

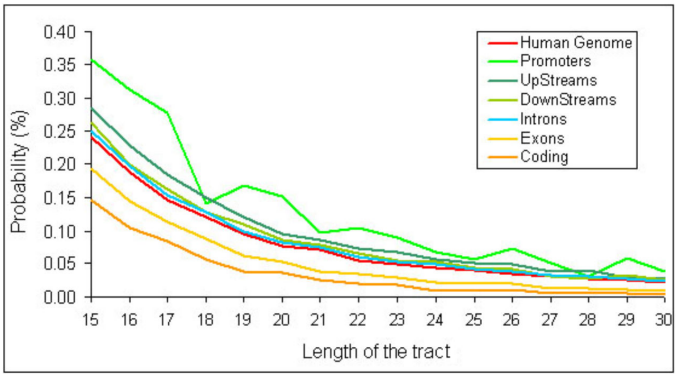

Figura 6.1. Distribució de probabilitats (segons la mida de la seqüència) de trobar un TTS en diferents regions del genoma humà.

L'estudi bioinformàtic a nivell genòmic de la localització de TTSs ha desvelat que el genoma humà és de fet un lloc especialment ric en seqüències dianes per la formació de tríplexs. Aquesta superpoblació de TTSs, molt per sobre del que un model d'atzar prediu, es concentra de forma diferent depenent de la regió genòmica. Tal com es mostra en la figura 6.1 On més rarament es torben els TTS és en les regions amb restriccions de seqüència per codificar proteïna (exons codificants). Les regions inter-gèniques i intra-gèniques no transcrites (introns) no mostren grans diferències respecte al general del genoma, mentre que les regions promotores acumulen de manera molt forta seqüències formadores de tríplex. Aquest resultat obre dos qüestions: i) es pot treure potencial biomèdic o biotecnològic a aquest fet?, i ii) per quina raó la natura ha col·locat TTS en una regió tan sensible com els promotors?.

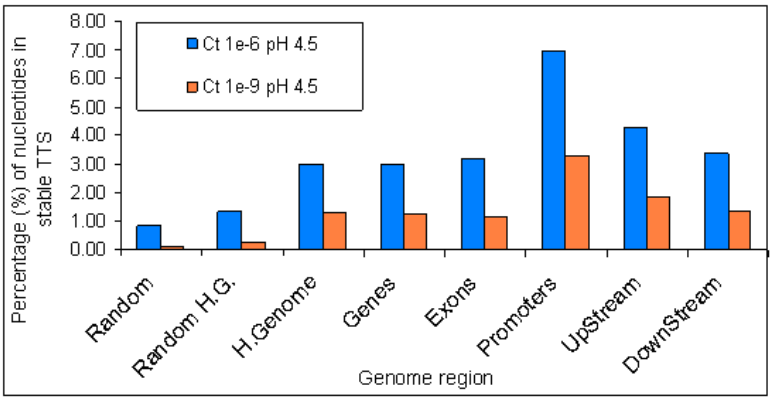

Figura 6.2 Estabilitat dels TTS trobats en regions humanes. Random i Random H.G. fan referència seqüències TTS generades aleatòriament considerant en l'últim cas la distribució de nucleòtids en el genoma humà. La estabilitat es mesura calculant el percentatge de nucleòtids en TTS estables a una temperatura superior a 50º (el càlcul s'ha fet a partir de dues concentracions de TFO:$1e^{-6}$ i $1e^{-9}$)

Per contestar a la primera pregunta hem d'analitzar la qualitat del TSS, que es mesura per l'especificitat i afinitat per la diana. L'avaluació bioinformàtica dels TTSs en el genoma humà ens revelat un resultat sorprenent. No només existeixen una concentració anormalment alta en dianes en els promotors, sinó que aquestes són més estables que les que es troben en altres regions (veure figura 6.2) i amb una longitud de TSS

moderada es pot aconseguir una especificitat total, especialment si només ens preocupen interaccions creuades amb altres regions promotores. La troballa de l'enriquiment de TTS de qualitat en regions promotores ha reforçat d'interès d'aquesta tècnica com a teràpia antigènica (Antony, Arimondo, Sun, & Pommier, 2004; Carbone et al., 2004; Coma, Noé, Eritja, & Ciudad, 2005, Duca et al., 2008).

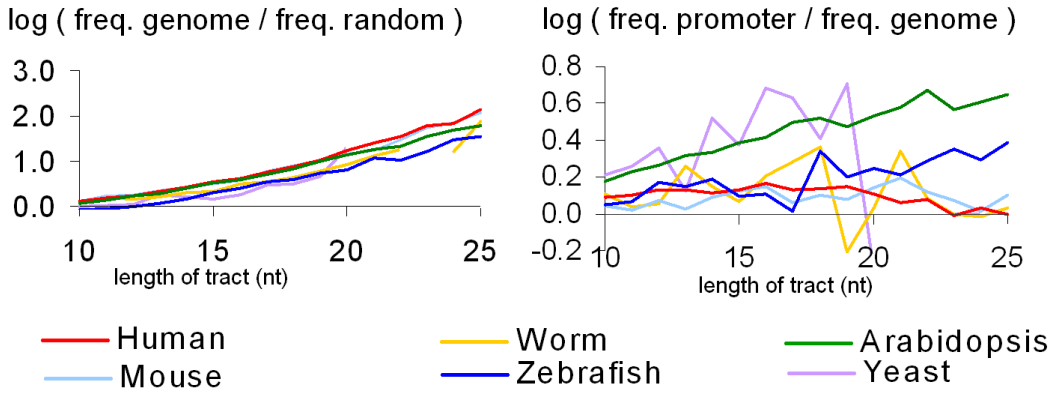

Figura 6.3. (Esquerra) Relació de la freqüència de TTSs en els genomes contra un model d'atzar. (Dreta) Relació entre freqüència de TTSs en els promotors contra la resta del genoma.

La segona qüestió (per que existeixen tants TTS en regions promotores) obre interessants possibilitats, ja que evidencies experimentals a partir de la immuno-detecció per anticossos anti-tríplex demostren que de fet els tríplex es produeixen de forma natural en el genoma humà (Ohno, Fukagawa, Lee, & Ikemura, 2002). Això suggereix la possibilitat de l'existència d'un mecanisme regulador basat en la formació de tríplex. Aquest mecanisme seria, com en cas dels RNAs d'interferència, possiblement antic i per tant hauria d'estar present altres organismes. Aquest punt es confirmat per un anàlisis en diferent especies que demostra que la super-població de TTS en promotors, lluny de ser una característica en genomes de mamífers també es dona en espècies inferiors com plantes o fins i tot la llevat (veure figura 6.3). L'indicador més clar però de que els TTS juguen un rol funcional en els mecanismes de regulació és el fet de que aquestes seqüències, no especialment conservades en posiciones llunyanes al TSS, mostren una tendència a estar-ho a mesura que s'aproximen a l'inici de transcripció (veure figura 6.4). En el promotor proximal el

perfil de les seqüències TTS i no TTS (aquestes últimes conegudes per actuar com segments de reconeixement per factors de transcripció) són idèntics.

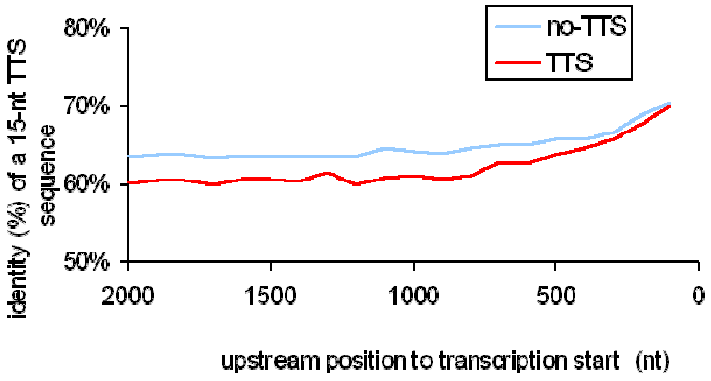

Figura 6.4 Conservació mitja dels TTS en regions properes a TSS entre home i ratolí. Els TTS estan normalment menys conservats en el genoma de però aquesta diferencia desapareix en posicions reguladores.

Els anàlisis de Gene Onthology (GO) revelen que entre els gens amb un TTS en la regió promotora estan mes representats els que codifiquen per factors de transcripció o proteïnes que s'uneixen a l'ADN i que a mes estan relacionats amb processos físics de regulació (veure figura 6.5). Un model que explicaria aquestes dades és el de la existència d'un mecanisme regulador *feed-back* per als gens (veure figura 6.6). Tot i que no s'han pogut trobar proves conclusives que demostrin (o refusin) l'existència d'aquest mecanisme, la nostra proposta ha obert el debat sobre l'existència de miRNA basats en tríplex (Hide, 2007). Treballs mes recents sobre en l'HIV han trobat indicis de que els miRNAs i els triplex-miRNAs podrien haver evolucionat per desactivar la infecció de l'HIV (i altres lentivirus) en alguns mamífers (O. Bagasra, 2006; O. Bagasra et al., 2006).

Malgrat el seu atractiu, la superpoblació de TTS en regions promotores no pot ser totalment explicada per l'existència d'un mecanisme residual d'autoregulació. Això re-obre la qüestió sobre la raó biològica de la presència de TTS en regions promotores, ja que al inserir-se disminueix el número de possibles dianes per factors de

transcripció i *a priori* disminueix el potencial de control en la expressió del gen. Potser, la resposta a aquesta pregunta pot estar no tant en la seqüència d'ADN sinó en les propietats físiques de la molècula que forma. Es a dir, coneixem que els TTS tenen unes restriccions en la seqüència que els identifiquen, i potser aquestes restriccions podrien dotar la doble hèlix de propietats especials essencials en la regulació dels gens.

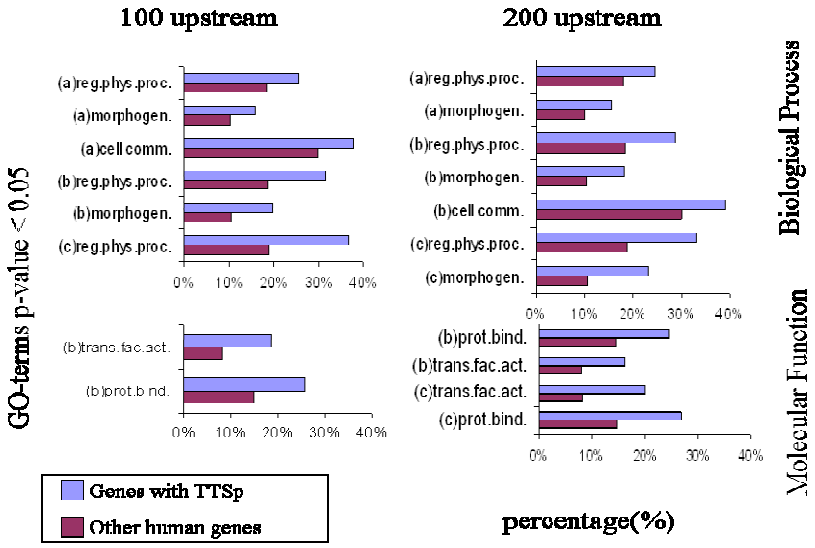

Figura 6.5 Enriquiment de termes GO en gens amb un TTS en el promotors. Els resultats són semblants agafant els 100 parells de bases (esquerra) o 200 abans que l'inici de transcripció. La mida del TTS mínim s'ha seleccionat en a) 15, b) 20 i 25) nucleòtids. La evidencia estadística s'ha realitzar mitjançant un test de Fisher amb el p-valor ajustat amb FDR.

La exploració de les propietats de TTS respecte a l'ADN genòmic s'ha centrat en quatre aspectes (1) la estabilitat, (2) la energia d'apilament, (3) la curvatura de la hèlix i (4) la flexibilitat de la seva estructura. Tal com s'aprecia en la figura 6.7 els TTS són lleugerament més corbats que la resta de l'ADN però sobretot són més rígids. Aquest resultat obre d'hipòtesis de que potser els TTS són elements separadors dels TFBS, que alhora ajuden a que els factors de transcripció es situïn correctament en l'espai tridimensional, afavorint la formació de complexes productius proteïna-proteïna.

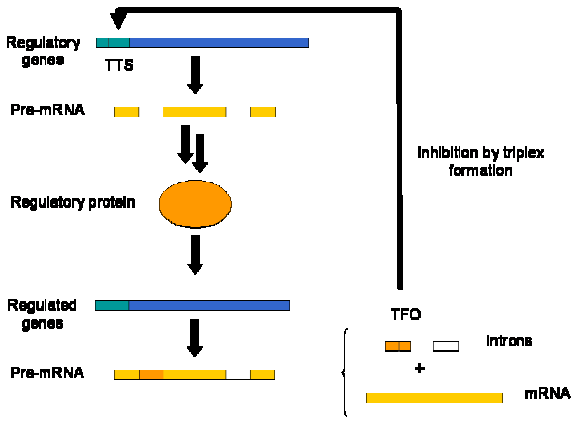

Figura 6.6 Diagrama d'un mecanisme regulador *feed-back* per a le control de la expressió d'un gen a través de la formació de tríplex. El TTS està en el promotor del gens reguladors i el TFO en l'intró del gens regulats.

L'aparent correlació entre TSS propietats físiques i regions reguladores ens va a obrir d'interès per entendre l'ADN des d'un punt de vista físic, intentat trobar sistemàticament alguna correlació entre propietats físiques i capacitat de regulació en el genoma dels mamífers i especialment al genoma humà.

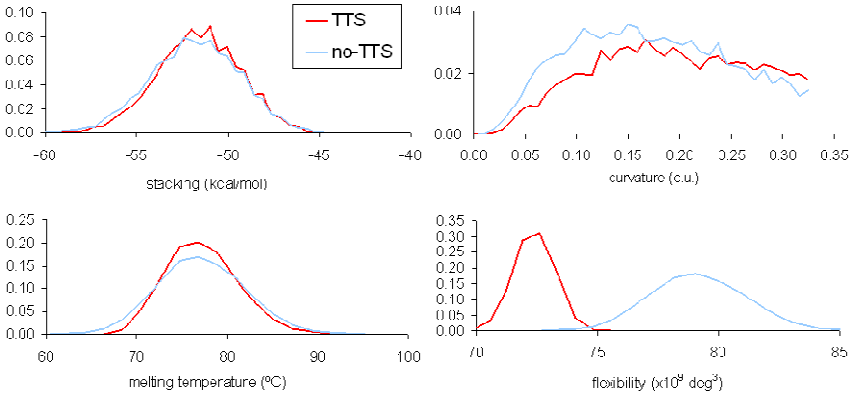

Figura 6.7 Estudi de 4 propietats físiques (estabilitat, energia d'apilament, curvatura y flexibilitat) entre seqüències dianes de tríplex i seqüències genòmiques.

6.2 EL ROL DE LES PROPIETATS FÍSIQUES DE L'ADN EN EL GENOMA HUMÀ

Alguns treballs han demostrat les característiques físiques úniques en promotors no humans (Pedersen, Jensen, Brunak, Staefeldt, & Ussery, 2000; Shpigelman, Trifonov, & Bolshoy, 1993). Això a motivat la idea de predir promotors a partir de propietats físiques, una tasca que no ha donat mals resultats en genomes procariotes (Kanhere & Bansal, 2005), però que no havia tingut massa èxit en genomes eucariota, especialment en organismes superiors com l'home que són molt mes complexos. Per exemple, en un treball recent Ohler et al. (Ohler, Nierman, Liao, & Rubin, 2001) en un intent per millorar la predicció *in silico* de TSS va refinar un predictor basat en seqüència incorporant propietats físiques de l'ADN, sense aconseguir un avanç important. No obstant, els nostres resultats en l'estudi de tríplex apuntaven clarament a que el perfil físic era un tret diferencial de les regions promotores en genomes superiors, pel que van decidir aprofundir en aquest tema.

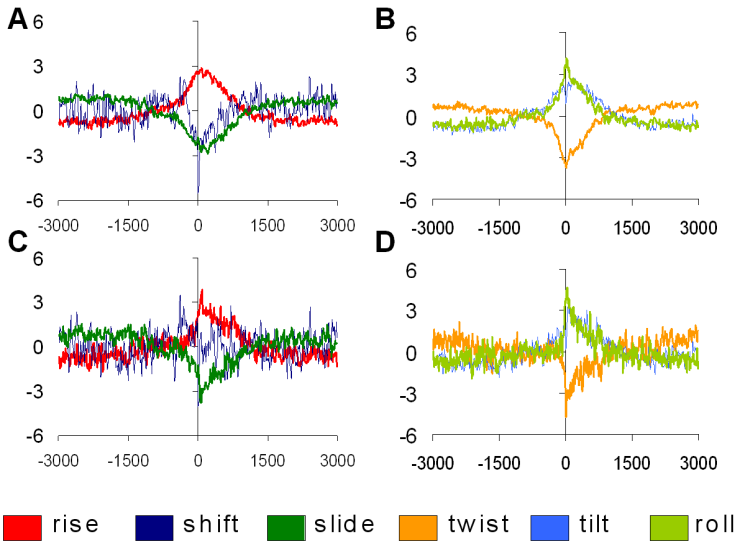

Figura 6.8 Flexibilitat en regions promotores. Valors promitjos per gens amb illa CpG (A i B) i gens sense illa CpG (C i D). L'inici de transcripció està localitzat a la posició +1. La flexibilitat anisotròpica s'ha descompost en 6 paràmetres helicoïdals.

Revisant els mètodes de predicció, el que es sospita és que la gran varietat de característiques físiques per descriure l'ADN i el fet de que potser poques d'elles siguin significatives en la predicció de promotors, pot introduir massa soroll per als algoritmes d'entrenament. Molts paràmetres poden estar fortament correlacionats entre si afegint informació innecessària al sistema, mentre d'altres es correlacionen fortament amb les illes CpG, sense contribuir especialment a la predicció de nous TSS, i tot això pot conduir a artefactes de sobre-entrenament que redueixen molt el poder predictiu dels algoritmes.

En aquesta tesis s'ha realitzat un estudi dels diferents tipus de paràmetres de l'ADN que ha determinat que només la flexibilitat mostra independència a la composició CpG dels promotors (veure figura 6.8). Aquests paràmetres de flexibilitat poden ser obtinguts de manera completament *ab initio* a partir de càlculs teòrics seguint models quasi-harmonics d'estudis de dinàmica molecular. El desenvolupament d'un *force-field* molt acurat per part del nostre grup, que ha permès l'estudi dinàmic de l'ADN en escales temporals del microsegon (Perez et al., 2007), ens va dotar de les eines necessàries per derivar un model global de flexibilitat basat exclusivament en paràmetres derivats de càlculs microscòpics. Aquest model és el que es va decidir provar com descriptor diferencial de l'existència de promotors.

En un intent de prioritzar les conclusions biològiques i de fugir del sobre-entrenament dels algoritmes predictius s'ha usat un discriminador lineal, la distància de Mahalanobis, per determinar el pes de cada paràmetre en la predicció. S'ha seguit escrupolosament el protocol d'EGASP per unificar criteris d'avaluació del poder predictiu del mètode. No s'han introduït consideracions de conservació inter-especie o el coneixement de l'estructura gènica per millorar el poder predictiu, ni tampoc s'ha usat cap algoritme predictor d'exons ni informació sobre seqüències diana de factors de transcripció.

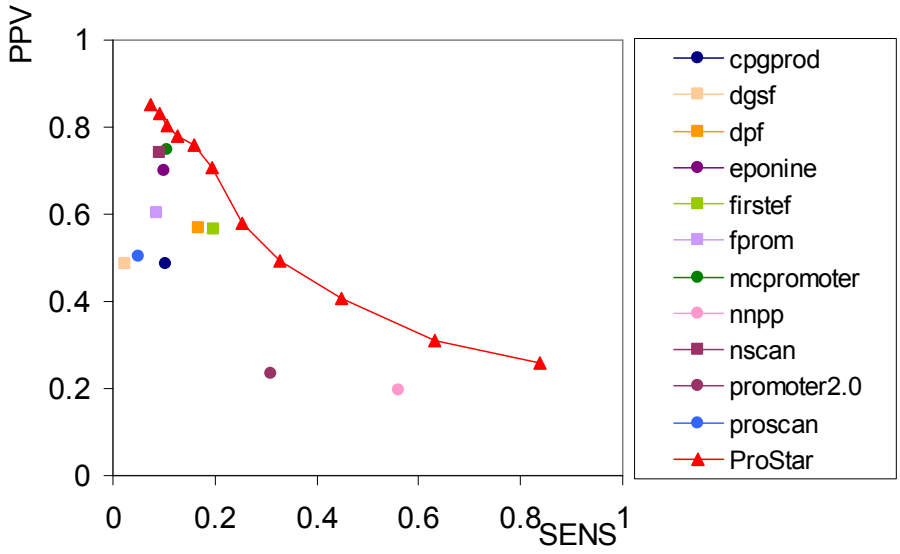

Figura 6.9 Sensitivitat i PPV de ProStar usant TSS CAGE dins de regions transcrites. El nostre mètode es mostra superior a qualsevol altre predictor avaluat.

El resultat d'aquest treball ha estat el programa ProStar que mostra un poder predictiu semblant als millors predictors de TSS (basats en conservació i predicció de gens) per a gens codificants en la regió Encode. Aquest avantatge es veu reduït quan els gens no codifiquen i per tant no tenen exons, on el nostre algoritme es demostra cada vegada mes (relativament) potent. El valor de ProStar es demostra plenament en la predicció d'inicis de transcripció no associats a cap gen, però que apareixen dins una regió transcrita, extrets a partir d'experiments CAGE (veure figura 6.9) on el nostre mètode apareix clarament com el més fiable. El resultats són especialment positius si tenim en compte que ProStar no ha esta entrenat per aquest tipus de TSS, sinó que ha tingut un entrenament genèric amb seqüències anotades per Havana. Els resultats de ProStar han esta avaluats també contra el genoma sencer i han mostrat una eficàcia similar a la regió Encode (veure figura 6.10) es a dir que el mètode tot i la seva extrema simplicitat te un fort poder predictiu i no sembla contaminat d'artefactes de sobre-entrenament.

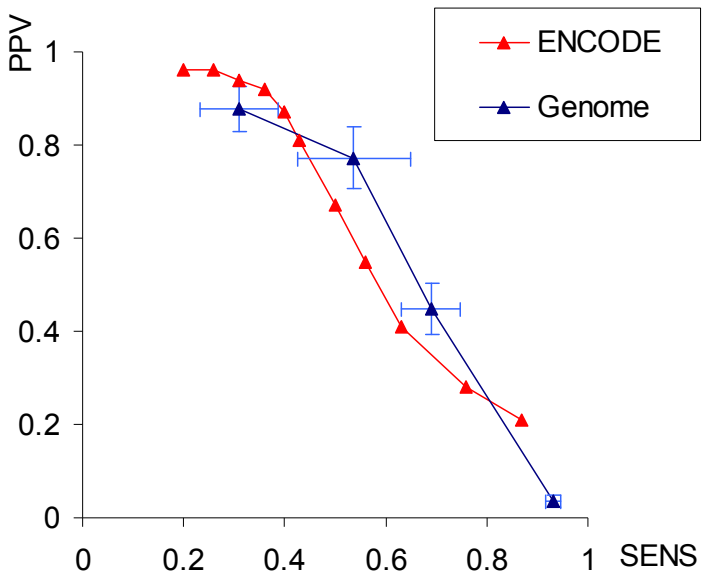

Figura 6.10 Comparació de ProStar en Encode i en la totalitat del genoma humà (usant Refseq). Les barres d'error corresponen a les diferències entre genomes.

ProStar no sub-classifica els promotors i fa una sola lectura per paràmetre. D'això se'n desprèn que existeix una propietat de flexibilitat universal en els promotors. Aquesta propietat física sembla que ha estat conservada, tot i que no a nivell de seqüència. Aquests resultats han esta confirmats de forma paral·lela per altres grups que han desenvolupat predictors basats en propietats físiques de l'ADN en procariotes (Singhal, Jayaram, Dixit, & Beveridge, 2008) i eucariotes (Abeel, Saeys, Bonnet, Rouzé, & Van de Peer, 2008). Actualment s'està treballant amb una versío més eficient de ProStar sense les restriccions inicials en l'entrenament i el tipus d'algoritme. ProStar 2 fa una pre-classificació del promotor i s'entrena segons el subtipus, obrint-se a nous paràmetres estructurals (Fenollosa et al. 2008, resultats no publicats) que permeten un augment en l'eficiència predictora. Malgrat la millora en els resultats ProStar 2 perd la simplicitat i l'elegància formal del nostre predictor original.

Els promotors no són els únics elements funcionals en el genoma que poden dependre de les propietats físiques, ja que molt possiblement tota l'estructura de la cromatina i per tant molts dels mecanismes de control epigenètic depenen de les propietats físiques

de l'ADN (Miele, Vaillant, d'Aubenton-Carafa, Thermes, & Grange, 2008). Per buscar elements físics inusuals i potencialment importants ocults en la seqüència de l'ADN hem desenvolupat la plataforma DNAlive (veure figura 6.11). DNAlive prediu 29 paràmetres estructurals de l'ADN que poden ser mostrats de forma combinada amb anotacions genòmiques (gens, exons, ...), facilitant a l'usuari la determinació de possibles regions de connexió entre funcionalitat (anotada per exemple en el Genome Browser) i propietats físiques inusuals.

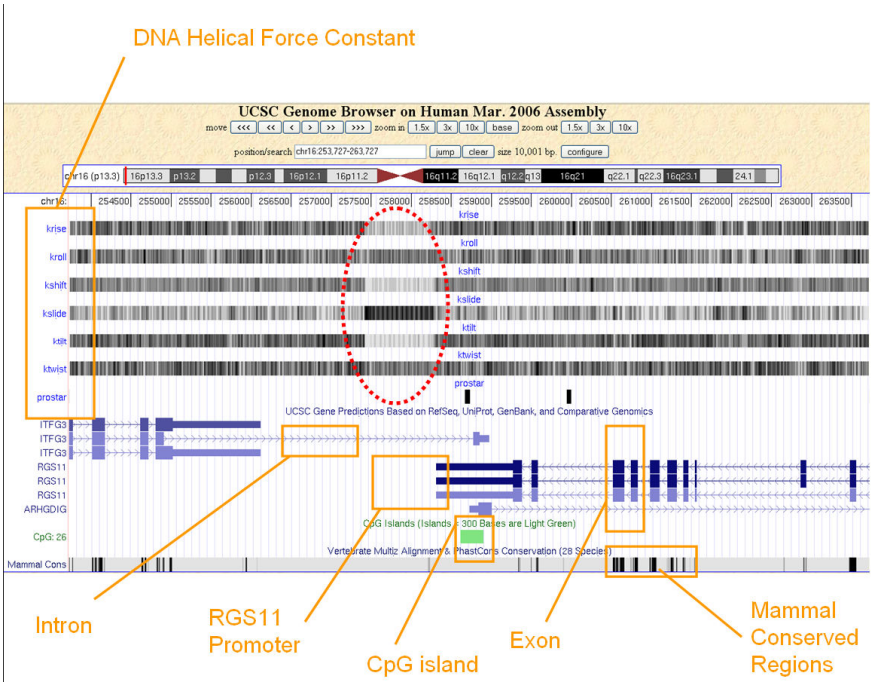

Figura 6.11. Combinació de propietats estructurals de l'ADN amb anotacions del GBD de la regió telomèrica del genoma humà 16p13.3 (Chr16:253:727-263:727). El promotor del gen RGS11 mostra una pertorbació de la flexibilitat clarament diferenciada (veure cercle discontinu vermell). En la mateixa posició co-existeix en sentit invers la regió 5'UTR de ITFG3.

Tot i que la representació en dos dimensions es compacta i poderosa existeix una manera més efectiva i realista de representar els paràmetres relacionats amb la estructura tridimensional de l'ADN: mostrar mapes genòmics en 3 dimensions. Aquest mapes permeten veure, per exemple, fins a quin punt dos regions aparentment distants de l'ADN poden arribar a interaccionar entre elles o si és possible la formació de

complexos de factors reguladors. A la figura 6.12 s'observa com DNAlive a integrat amb èxit un algoritme que prediu la estructura tridimensional d'equilibri de l'ADN (tenint en compte efectes de seqüència), incorpora estructures de proteïnes i pot usar una codi de colors per anotar qualsevol tipus d'anotació. La representació precisa de promotors està però subjecta a que: i) l'usuari conegui en profunditat els llocs d'unió del TF (ja que la predicció *in silico* d'aquest elements no es mostra massa acurada) i ii) que existeixi la estructura tridimensional de cada un dels TF amb el segment d'ADN on s'uneixi. El repte d'aqueta plataforma en el futur és el de millorar la predicció de llocs d'unió a proteïnes, ser capaç de predir com s'uneix un TF a l'ADN i de predir i representar interaccions entre proteïnes.

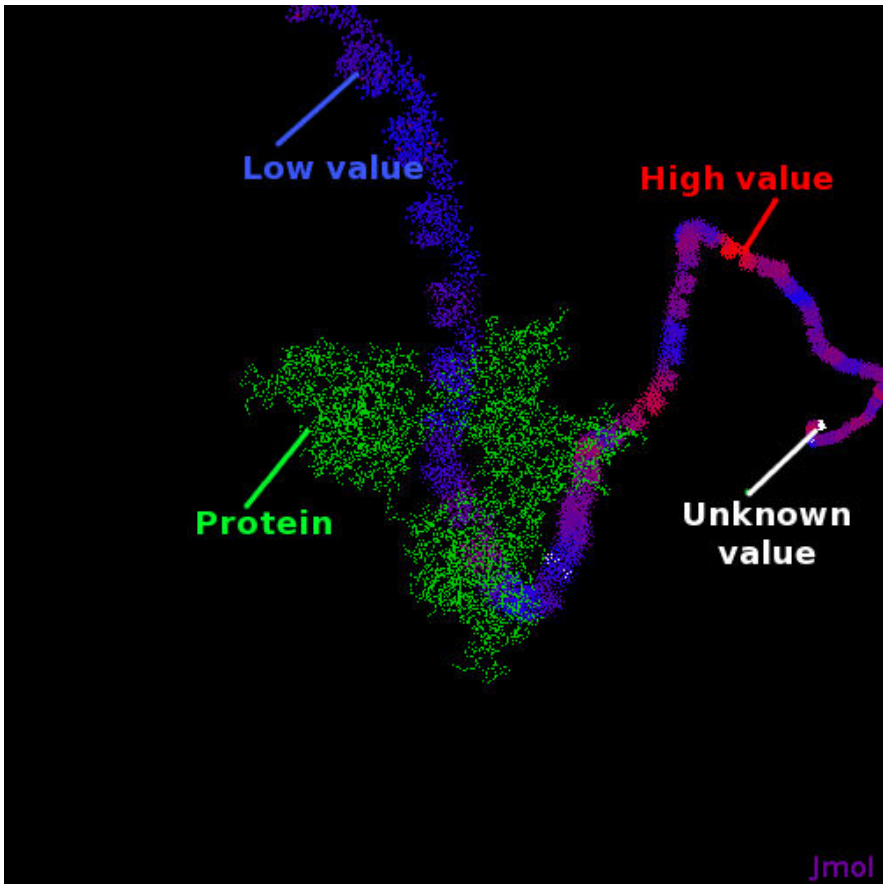

Figura 6.12 Fotograma de la representació 3D d'un motiu genòmic amb proteïnes. El degradat de color representa una propietat de flexibilitat. DNAlive permet la rotació i zoom de la imatge en

temps real.

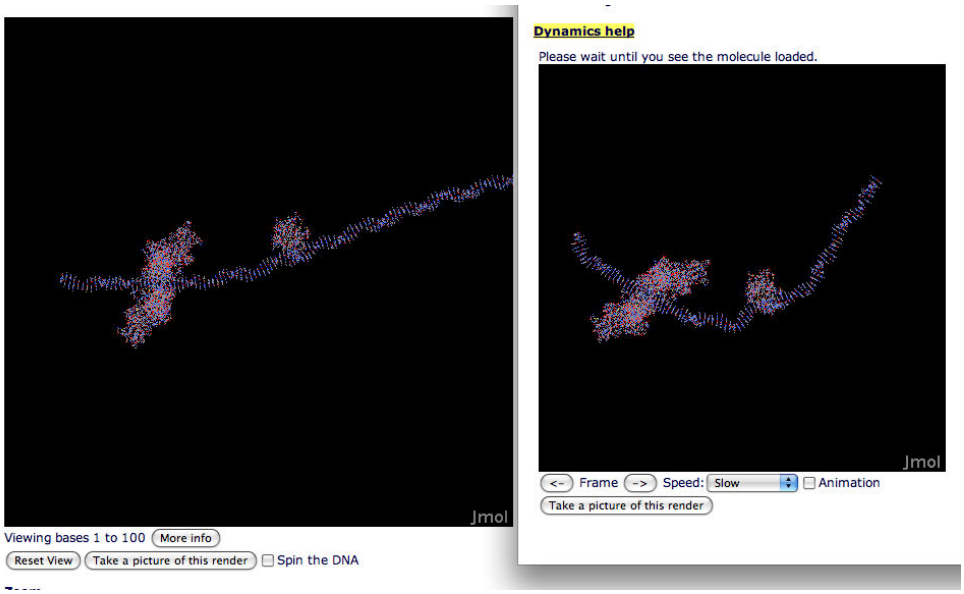

Figura 6.13 Dinàmica d'un segment d'ADN unit a dues proteïnes. La dinàmica (dreta) es calcula a partir d'un recorregut de Montecarlo de la estructura en equilibri (esquerra) calculada per DNAlive.

L'element més avançat de DNAlive es la capacitat d'incorporar efectes de flexibilitat dinàmica a escala genòmica. Per això el programa incorpora un algoritme de Montecarlo que fa servir una descripció macroscòpica de l'ADN, a on sense perdre resolució a nivell dels parells de bases aconseguim similar (amb una qualitat molt acceptable) el comportament dinàmic de grans fragments de ADN (5,000 parells de bases) d'una manera extremadament ràpida (veure figura 6.13). Això permet la extensió del marc d'aplicació de DNAlive i estudiar la interacció distant entre proteïnes reguladores o regions llunyanes pot ser o no possible, fora de la situació d'equilibri en funció de la flexibilitat (depenent de seqüència) de l'ADN. Actualment està en desenvolupament una versió més sofisticada de l'algoritme de càlcul de dinàmiques genòmiques de DNAlive. Primer es vol incorporar la producció d'una dinàmica essencial que sintetitzi els moviments mes rellevants de l'ADN. Segon es treballa amb un model que sigui mes realista davant les possibles col·lisions de l'ADN (Pérez et

al. 2008, resultats no publicats). Finalment, estem incorporant efectes epigenètics relacionats amb el canvi en l'estructura covalent de l'ADN. Amb el que serà possible la simulació de segments de pràcticament qualsevol longitud en qualsevol estat de compactació i en qualsevol estat de modificació epigenètica.

L'empaquetament dels algoritmes de DNAlive en serveis-web publicats en l'Institut Nacional de Bioinformàtica ofereix a la comunitat científica un canal flexible i obert per al seu ús (veure figura 6.14). La integració de DNAlive amb MODEL (Rueda et al., 2007) la principal plataforma de dinàmiques de macromolècules és una opció que s'està avaluant.

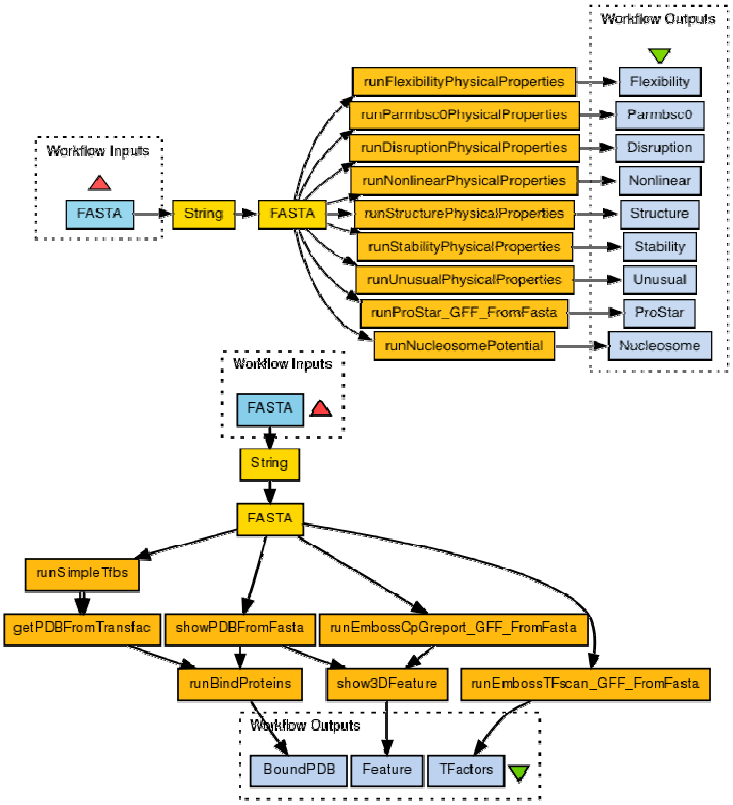

Figura 6.14 Diagrama de dos fluxes de treball construïts a partir dels serveis-web de DNAlive i disponibles en l'INB.