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

4 MÈTODES

Els mètodes usats en la elaboració d'aquesta tesis s'engloben en l'àmbit de la biologia computacional. El treball presentat s'ha basat en el desenvolupament de programes informàtics per explorar i validar les hipòtesi de treball formulades. Les eines bioinformàtiques han estat normalment una via per avançar la recerca i en última instància un fi en sí mateixes especialment en el cas dels programes desenvolupats DNAlive i ProStar (veure capítol Resultats).

El treball bioinformàtic ha representat no només el desenvolupament de software propi sinó també la incorporació i millora d'eines bioinformàtiques ja existents i que estan descrites en detall en la següent secció. La segona part d'aquest capítol fa referència a les bases de dades biològiques usades el treball d'investigació.

4.1 EINES BIOINFORMÀTIQES

La bioinformàtica i la biologia computacional integren tècniques matemàtiques, informàtiques estadístiques i les pròpies de la química computacional per estudiar els problemes biològics. El principi d'aquestes tècniques és usar el poder de càlcul de les computadores modernes per solucionar problemes d'impacte biològics i que per la seva complexitat ultrapassen la capacitat humana. Els esforços principals de la bioinformàtica inclouen l'alineament de seqüències, la cerca i anotacions de gens i dels seus transcrits, la determinació dels mecanismes de regulació gènica, la compilació de genomes, l'alineament i predicció d'estructures de proteïnes, l'estudi de la interacció entre macromolècules, l'estudi de la dinàmica de les macromolècules, el estudi de la biologia dels sistemes, el disseny de lligands i la descripció sintètica de la biologia.

Els mètodes bioinformàtics usats i exposats en aquesta memòria s'han dividit en tres grans blocs: i) predicció de les propietats fisicoquímiques de la molècula d'ADN, ii) la predicció d'anotacions en el genoma (com els gens, els promotors, els llocs d'unió dels factors de transcripció o les posicions ocupades pels nucleosomes) i finalment iii) les eines per l'estudi de la dinàmica de l'ADN.

4.1.1 Mètodes de predicció de propietats físiques de l'ADN

Un nombre important de mètodes de predicció del descriptors físics intrínsecs a l'ADN es poden trobar en la bibliografia (Florquin, Y. Saeys, Degroeve, Rouze, & Y. Van de Peer, 2005; J Ramon Goñi, Fenollosa, Pérez, Torrents, & Modesto Orozco, 2008; J. V. Ponomarenko et al., 1999). De forma esperada molts d'aquests paràmetres estan altament correlacionats i el repte és determinar quin paper real poden jugar en el funcionament dels mecanismes reguladors del gens (Ohler, Liao, Niemann, & G. M. Rubin, 2002).

Els descriptors fisicoquímics de l'ADN que poden ser predits experimentalment o computacionalment són normalment dependents de la longitud i la seqüència de l'ADN. En molts paràmetres, els efectes d'interacció entre parells de bases no-consecutius a la seqüència són negligibles o, si mes no, la seva contribució és poc rellevant. Molts dels mètodes explorats en aquest capítol estan basats en una taula de correspondències entre un dinucleòtid (o trinucleòtid, etc.) i el valor derivat de la seva contribució en un paràmetre estructural (veure figura 4.1).

Un mètode generalitzat per determinar la contribució lineal de cada dinucleòtid a una propietat física de l'ADN, a partir de resultats experimentals, és l'encaix de paràmetres pel veí-proximal (*nearest-neighbourd*). Per exemple, el cas del mètode derivat per Santalucia (Santalucia, 1998) per estimar la temperatura de fusió d'una seqüència curta d'ADN, on els paràmetres assignats a cada dinucleòtid estan derivats a partir del càlcul experimental de la temperatura de 108 oligonucleòtids (veure figura 4.1). La taula de correspondència entre dinucleòtid i paràmetre es pot resoldre per una regressió múltiple lineal (J. Cohen, P. Cohen, West, & Aiken, 2002). En alguns casos el valor final de la predicció cal ser ajustat per algun tipus d'elements no-lineal. Per exemple, en el càlcul de la temperatura de fusió factors com els nucleòtids terminals o la simetria de la seqüència tenen una repercussió en l'efecte d'obertura de la doble hèlix i cal que siguin considerats en el càlcul final (Santalucia, 1998).

Dinucleotide	ΔH° kcal/mol	ΔS° cal/k·mol
AA/TT	-7.9	-22.2
AT/TA	-7.2	-20.4
TA/AT	-7.2	-21.3
CA/GT	-8.5	-22.7
GT/CA	-8.4	-22.4
CT/GA	-7.8	-21
GA/CT	-8.2	-22.2
CG/GC	-10.6	-27.2
GC/CG	-9.8	-24.4
GG/CC	-8	-19.9
Init. w/term. G·C	0.1	-2.8
Init. w/term. A·T	2.3	4.1
Symmetry correction	0	-1.4

$$\Delta G^\circ_T = \Delta H^\circ - T\Delta S^\circ.$$

```
      ↓  ↓  ↓
5′ C-G-T-T-G-A 3′
   * * * * * *
3′ G-C-A-A-C-T 5′
      ↑  ↑
```

ΔG°_{37}(pred.) = ΔG°(CG/GC) + ΔG°(GT/CA) + ΔG°(TT/AA)
+ ΔG°(TG/AC)+ ΔG°(GA/CT) + ΔG°(init.)
= -2.17 - 1.44 - 1.00 - 1.45 - 1.30 + 0.98 + 1.03

ΔG°_{37}(pred.) = -5.35 kcal/mol

ΔG°_{37}(obs.) = -5.20 kcal/mol

Figura 4.1 Aplicació del mètode de predicció de la temperatura de fusió d'un oligo (veure equació de la figura) a partir de la taula de paràmetres veins-propers derivats experimentalment (Santalucia, 1998). Cada fletxa està entre un dinucleòtid NN. En aquest exemple, el dúplex CGTTGA·TCAACG no és auto-complementari i per tant la ΔG°(sym) val zero.

Una estratègia diferent per determinar la contribució lineal de dinucleòtids en les propietats físiques de l'ADN és l'ús de tècniques computacionals per simular la molècula d'ADN (veure seccions posteriors). És el cas, del treball de Sponer *et al.* en la derivació de les contribucions de les energies d'apilament i pont d'hidrògen a partir de càlculs mecanico-quàntics d'alta qualitat pels 10 parells de bases (J. Sponer, Gabb, Leszczynski, & P. Hobza, 1997).

Sigui el cas que els paràmetres es derivin a partir de dades experimentals o de càlculs de simulació, els descriptors es poden classificar en quatre grans grups.

- Paràmetres fisicoquímics de l'ADN relacionats amb la estabilitat de la doble hèlix.

- Paràmetres estructurals de la doble hèlix que determinen la seva estructura tridimensional.

- Flexibilitat de la doble hèlix, ja siguin intrínseca o induïda de forma externa per una proteïna.

- Regions afins a formes no canòniques de l'ADN, tals com el Z-DNA, els tríplexes o els tètraplexes.

4.1.1.1 Paràmetres fisicoquímics de l'ADN

Conjunt de descriptors que informen sobre estabilitat de l'ADN genòmic i la seva resistència a la desnaturalització parcial o total.

4.1.1.1.1 Energia lliure de la hèlix

La energia lliure de la hèlix d'ADN representa la estabilitat de la molècula i està estretament relacionada amb la temperatura de fusió. Els paràmetres d'energia associats a cada dinucleòtid emprats a aquesta tesis han estat derivats a partir de valors experimentals han estat derivats a partir d'estudis calomètrics de desenes de seqüències d'ADN (Breslauer, Frank, Blöcker, & Marky, 1986; Sugimoto, Nakano, Yoneyama, & Honda, 1996).

Usant una altra estratègia els paràmetres d'apilament han estat derivats a partir resultats teòrics de simulació. Així s'han fet servir resultats derivats de camps de força simples usant les geometries en equilibri (Ornstein, Rein, Breen, & Macleroy, 1978), i més recentment, paràmetres determinats *ab initio* a partir de càlculs quàtico-químics d'alta qualitat (J. Sponer et al., 1997).

4.1.1.1.2 Temperatura de fusió de l'ADN

Existeix un gran repertori de mètodes de predicció de paràmetres relacionats amb la temperatura de fusió de l'ADN (R D Blake & S G Delcourt, 1998; Gotoh & Tagashira, 1981; SantaLucia, 1998) la majoria dels quals produeix resultats qualitativament similars. Aquests mètodes mostren una sorprenent precisió en seqüències petites d'ADN (10~30 nucleòtids), encara que són més imprecisos per seqüències llargues (veure següent apartat). En la majoria dels casos el model incorpora elements no

linealment relacionats amb els dinucleòtids com ara la simetria de la seqüència, junt a termes lineal relacionats amb la hipòtesis dels veïns mes propers.

4.1.1.1.3 Càlcul de la dinàmica de fusió de fragments llargs d'ADN

La desnaturalització d'un fragment petit de ADN és un procés de dos estats, que és fàcil de predir i representar, però en una fibra llarga el procés es multi-etapa ja que, el desplegament es produeix probablement de manera simultània a diferents punts, generant bombolles de desplegament que influencien la estabilitat d'altres regions, fent per tant invalida la aproximació lineal implícita als mètodes comentats més amunt. Per resoldre el problema, Peyrard i els seus col·laboradors (van Erp, Cuesta-Lopez, Hagmann, & Peyrard, 2005) han dissenyat i ajustat un model molt més complex que té en compte els efectes no-lineals de dissociació entre les dues cadenes (veure capítol Estructura de l'ADN) per poder realitzar una predicció efectiva de seqüències genòmiques (veure figura 4.2). El model està derivat del les simulacions dinàmiques de Peyrard-Bishop-Dauxois (Dauxois, Peyrard, & Bishop, 1993) que intenta reproduir les corbes de desnaturalització de petits segments d'ADN, el que a'ha sugerit pot ser útil per localitzar llocs d'inici de transcripció (veure Figure 4.2). El cost computacional associat a aquest càlcul és però molt elevat i escala amb la mida del segment d'ADN que es considera.

4.1.1.2 Predicció de paràmetres estructurals de la doble hèlix

A partir d'estructures de cristalls de molècules d'ADN es poden derivar tant els paràmetres de d'angle i direcció de la curvatura (A Bolshoy, McNamara, Harrington, & E N Trifonov, 1991; Goodsell & R E Dickerson, 1994) com els paràmetres helicoïdals de la hèlix (A A Gorin, V B Zhurkin, & W K Olson, 1995; el Hassan & Calladine, 1996). Repetint aquest procés per moltes seqüències es possible derivar paràmetres promigs a nivell de veïns (2-mer), tripletes (3-mer) o quartets (4-mer).

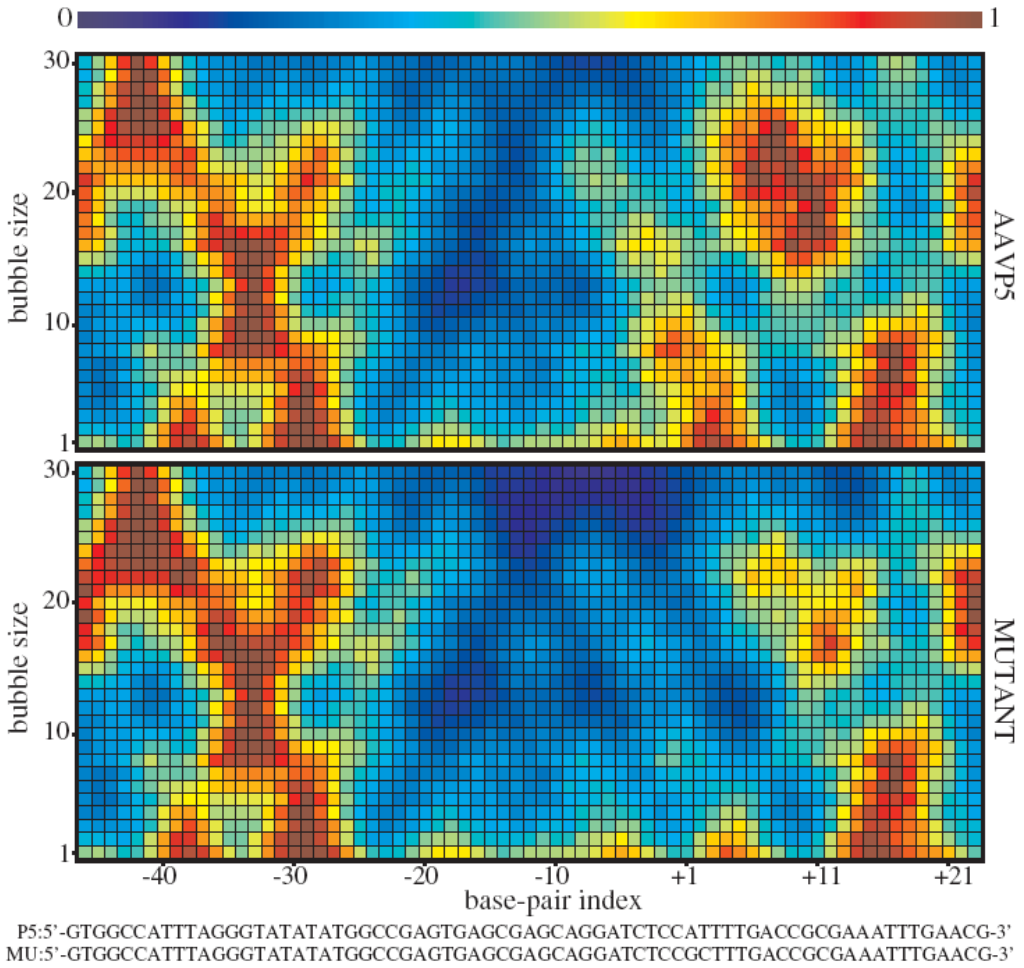

Figura 4.2. Càlcul de la probabilitat de formació de bombolles (van Erp et al., 2005) en el promotor del AAVP5. La seqüència calculada està en el rang de -46 a +23 on el TSS es la +1. El gràfic superior fa referència a la seqüència original, mentre el segon mostra l'efecte d'una mutació a la posició (+1,+2).

Aquest paràmetres estructurals són molt útils per descriure una fibra de B-DNA en estat d'equilibri, però no podem oblidar que en realitat l'ADN pot patir canvis estructurals molt importants per la intervenció de factors externs, com per exemple, la unió d'una proteïnes, canvis en el pH, el grau d'hidratació, la temperatura, el grau d'humitat, la presència de petits lligands, etc. Aquest canvis són molts cops dramàtics tal com s'ha vist en estudis dels paràmetres d'estructures de cristalls amb complexos ADN-proteïna (W. K. Olson, A. A. Gorin, X. Lu, L. M. Hock, & V. B. Zhurkin, 1998). En conclusió,

és necessari complementar la informació estructural sobre l'ADN afegint-li descriptors que informin sobre la seva capacitat de deformació (veure figura 4.3).

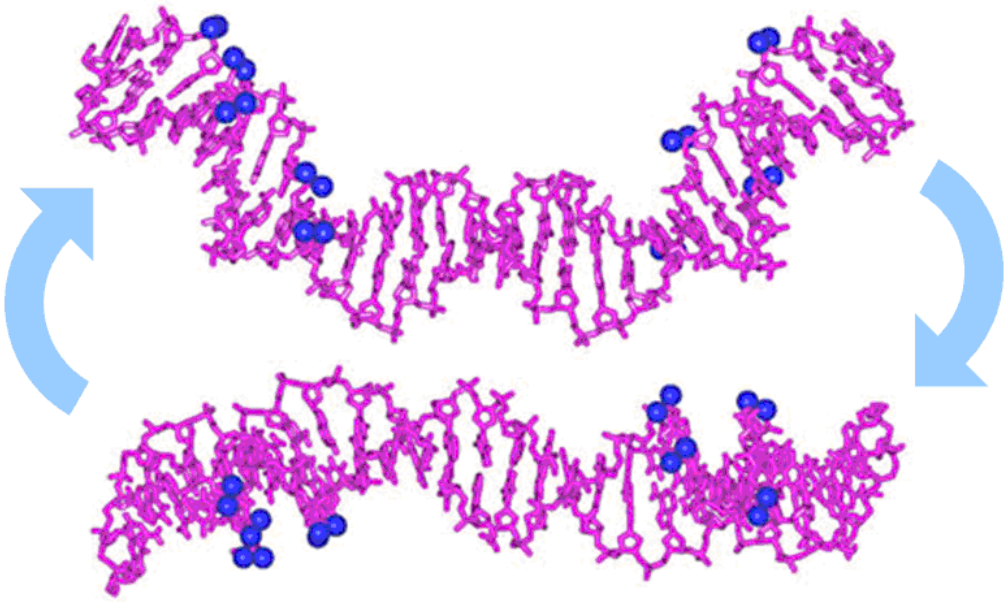

Figura 4.3 La estructura tridimensional de la doble hèlix és una propietat dinàmica de l'ADN. Existeixen unes conformacions més estables que representen majoritàriament la forma de la molècula, però aquest varia en el temps cap a formes menys favorables energèticament, especialment en els casos de la intervenció d'una proteïna externa.

4.1.1.3 Predicció de la flexibilitat de la doble hèlix

S'ha suggerit l'ús del potencial de la deformació harmònica (Brukner, Sánchez, Suck, & Pongor, 1995; W K Olson, A A Gorin, X J Lu, L M Hock, & V B Zhurkin, 1998a) per a cada parell de bases depenent de cada paràmetre helicoïdal (Roll, Twist, Shift, Slide i Rise) per tal de descriure la flexibilitat a prop de l'equilibri de la fibra d'ADN. Aquest model, tot i la seva simplicitat descriu raonablement be a la flexibilitat de l'ADN i permet introduir efectes de seqüència que són sovint ignorats en models més simples como el *rod-model*. és de destacar que aquesta dependència per seqüència de

la flexibilitat és molt important, ja que per exemple, el cost energètic de desenroscar el dinucleòtid CG és la meitat que el del pas AC (J Ramon Goñi et al., 2007). Existeixen passos universalment flexibles, com el dinucleòtid AT o generalment rígids com el AC, però en general la flexibilitat o rigidesa diferencial d'un pas depèn del tipus de pertorbació introduïda.

Wilma Olson y col·laboradors van desenvolupar un mètode simple en el que es determinaven 6 paràmetres de flexibilitat per cada pas (2-mer), corresponents a les deformacions de rise, slide, shift, tilt, twist and roll. Els paràmetres es derivaven a partir de l'anàlisis de la variació dels diferents parametres helicoïdals (rise, slide, sfith, tilt, twist and roll) per cada un dels deu passos ùnics (d(AA), d(AG),d(AT),d(AC),d(GG), d(GA),d(GC),d(TA),d(TG) and d(CG)) en estructures cristal·lines (W K Olson, A A Gorin, X J Lu, L M Hock, & V B Zhurkin, 1998b). Invertint la matriu de covariança per cada pas i applicant l'ecuació d'Einstein es possible derivar una matriu de rigidesa (stifnness matrix) que informa sobre la resistencia d'un pas a patir deformacions harmòniques. El problema d'aquests paràmetres és que tenen una forta desviació per la sobre representació de seqüències riques en G/C a les estructures cristal·lines (Beveridge et al., 2004). Per evitar aquest efecte, Lankas *et. al* van usar les constants de deformació harmònica derivades de simulacions en dinàmiques moleculars (MD) (Lankas, Jirí Sponer, Langowski, & Thomas E Cheatham, 2003). Seguin aquests mètodes, a partir de les MD de petites seqüències d'ADN (que contenen tots els passos de dinucleòtids possibles) les constants de força helicoïdals es deriven a partir de la inversió de la matriu de co-variança en l'espai helicoïdal i assumint que totes les oscil·lacions són harmòniques (J Ramon Goñi et al., 2007), simulant en això el procediment propossat per W.Olson, pero sense tenir ara problemes de desbalanç en la representació dels diferent passos. El problema de la tècnica de Lankas es que es molt depenent del camp de força. L'ús d'un nou camp de força (anomenat PARMBSC0; (Perez et al., 2007)) ha permès un refinament de les constants de força a partir de trajectòries més llargues i estables, millorant molt la precissió de les constants derivades.

4.1.1.4 Predicció de regions afins a formes no canòniques

Tal com s'ha descrit en el capítol anterior, la forma principal d'ADN és el B-ADN, a la qual fan referència els mètodes predictors anteriorment esmentats. Existeixen, però, altres conformacions minoritàries possibles per la gran llibertat d'aparellaments físics entre les bases. Aquestes conformacions, que es troben *in vivo* en la cèl·lula i poden tenir un important paper funcional, solen tenir motius de seqüència favorables o restriccions en la seva composició. A partir de l'estudi experimental de l'estabilitat d'aquestes molècules se'n poden derivar mètodes que prediguin la preferència en el genoma per la formació de conformacions no canòniques d'ADN. Així, la tendència de l'ADN per adoptar la forma A-ADN (molt comú en complexes amb proteïnes) ha estat molt estudiada derivant-se paràmetres a nivell de veïns de seqüència a partir de mesures experimentals (Ivanov & Minchenkova). De manera similar (Ho, G. W. Zhou, & L. B. Clark, 1990) s'han derivat descriptors de la tendència del DNA a formar la estructura tipus Z. Respecte a les formes triple-cadena existeixen estudis pels tríplex paral·lels que han permès derivar paràmetres específics d'estabilitat per seqüència, concentració de TFO (triplex forming oligonucleotide) i pH (R. W. Roberts & Crothers, 1996). Per tríplex anti-paral·lels la predicció de la estabilitat podria simplificar-se ja que aquests no són dependents del PH (Jaumot, Aviña, R Eritja, Tauler, & Gargallo, 2003), pero malauradament no tenim encara una escala fiable. Per últim estudis de calorimetria han permès derivar paràmetres per predir l'estabilitat de G-DNA o tétraplex basat en el número de tètrades de guanines apilades, la llargada i la composició de la seqüència (Huppert & Balasubramanian, 2005).

4.1.2 Mètodes de predicció d'anotacions en genomes

L'anotació del genoma és per definició el procés d'associar una informació biològica a les seqüències. La identificació d'aquesta informació és una tasca que pot realitzar-se manualment, requerint tota la habilitat d'un expert humà o, de forma automàtica gràcies a l'ús d'eines d'anàlisis computacional de la seqüència. Existeixen diferents tipus d'anotacions, les més rellevants per la present tesis són la predicció dels gens, la predicció de regions promotores o TSS, la predicció de llocs d'unió de TF i la predicció de seqüències posicionadores de nucleosomes.

4.1.2.1 Predicció de gens

És un dels elements claus en l'anotació d'un genoma i malgrat tot l'esforç realitzat és encara un tema obert, especialment pel que fa referència a gens no codificants o d'evolució molt ràpida. Una aproximació simple al problema seria detectar una pauta oberta de lectura (en anglès *open reading frame*; ORF), determinar el codó inicial (ATG) i el final (TAA, TAG o TGA). L'aproximació és sovint útil a procariotes, però malauradament, de seguida es va veure que en organismes eucariotes, més evolucionats era poc fiable degut a diferents raons:

- En genomes eucariotes els gens es poden sobreposar uns als altres (inclús es pot donar el cas que un gen estigui dins un altre gen)

- Les regions intròniques que interrompen els exons no tenen senyals clares per les que puguin ser reconegudes

- Un mateix gen pot disposar de múltiples variants (múltiples inicis, múltiples finals, ...)

La problemàtica per determinar el posicionament dels gens ha conduit al desenvolupament d'un gran nombre de mètodes de predicció, que fan servir una o varies de les següents estratègies de localització:

- Aproximació *ab initio:* intenten predir gens basant-se en propietats estadístiques de la seqüència d'ADNs que apareixen en gens coneguts. Algoritmes d'aquest tipus estan incorporats a programes com Genscan (Burge & Karlin, 1997), Geneid (R Guigó, 1998), Genie (Martin G. Reese, David Kulp, Tammana, & David Haussler, 2000) o Fgneh (Solovyev & Salamov, 1997).

- Aproximacions basades en l'homologia. La seqüència d'ADN candidata a ser anotada com gen es compara amb una proteïna coneguda o amb una seqüència (total o parcial) d'ARNm. Programes com el desenvolupat per Gelfand et al. (Gelfand, Mironov, & Pevzner, 1996), GeneWise (Ewan Birney, Michele

Clamp, & Richard Durbin, 2004) o el Genomethreader (Gremme, Brendel, Sparks, & Kurtz, 2005) es basen en aquesta estratègia.

- Anàlisis comparatiu. Es basa en l'assumpció que els exons, estan mes ben conservats en altres genomes que la resta d'ADN. Basats en aquesta idea existeixen entre d'altres programes CEM (Bafna & Huson, 2000) o el Twinscan (I Korf, P Flicek, D Duan, & M R Brent, 2001a).

4.1.2.1.1 Mètodes *ab initio*

Els mètodes *ab initio* afronten el problema de la predicció des de la perspectiva més difícil ja que, ignoren l'informació pre-existent sobre proteïnes i gens anotats, però son útils perquè (Michael R Brent & Roderic Guigó, 2004):

- Són més fàcils d'entrenar i són molt més ràpids que els mètodes que involucren informació externa

- Les senyals de seqüència funcionals en un gen (llocs d'estroncament) o la pròpia composició específica de les regions codificants són elements intrínsecs de la seqüència d'ADN.

- Qualsevol altre tipus de predicció requereix el coneixement previ d'altres elements com per exemple, la seqüència d'aminoàcids de les seves proteïnes (predicció d'homologia) o dels genomes d'espècies properes (predicció basada en la conservació).

Les prediccions *ab initio* treballen principalment amb característiques de l'ADN, entenent en aquest cas com a característica qualsevol sub-seqüència d'ADN amb significança biològica. El consens general es classificar en dos tipus aquestes característiques:

- Característica de senyal: característiques de longitud fixada i generalment curta; com per exemple, el codó inicial

- Característica de regions amb contingut: característiques de longitud variable; com per exemple, els exons

Així doncs els programes predictors *ab initio* entren dos tipus de sensors:

- Sensors de senyal: escanegen una seqüència d'ADN, desplaçant una finestra o marc de mida fixada, identificant posicions on és versemblant la presència d'una senyal.

- Sensors de contingut: sensors usats per assignar una puntuació a regions que separen dues senyals. Per exemple, un sensor d'introns puntua la regió entre la senyal d'inici d'estroncament i final d'estroncament.

A = 31% T = 28% C = 21% G = 20%	A = 18% T = 32% C = 24% G = 26%	**A** 100%	**T** 100%	**G** 100%	A = 19% T = 20% C = 29% G = 32%	A = 24% T = 18% C = 26% G = 32%
C	T	**A**	**T**	**G**	A	C
0.21	0.32	**1.0**	**1.0**	**1.0**	0.19	0.26

$$P(\mathrm{CT}\underline{\mathbf{ATG}}\mathrm{AC}|q_{start\_codon})$$
$$= 0.21 \times 0.32 \times 1.0 \times 1.0 \times 1.0 \times 0.19 \times 0.26$$
$$= 0.00331968$$

Figura 4.4 Construcció i ús d'una matriu de pesos per predir la senyal d'inici de traducció. La matrius es construeix entrenant una cadena de Markov d'ordre 0, que representa la probabilitat de trobar cada una de les bases en cada posició relativa a la senyal ATG. Per avaluar la seqüència només cal aplicar la taula (part inferior) i calcular-ne la probabilitat.

Les senyals típiques de les característiques dels gens eucariotes són i) ATG pel codó inicial (o inici de traducció), ii) TAG, TGA o TAA pel codó final, iii) GT donadora

d'estroncament, iv) AG per acceptador d'estroncament i v) AATAAA o ATTAAA per la senyal de poli-adenilació.

En molts sistemes el reconeixement de senyal es basa en models probabilístics. Per exemple, el model més comú per reconèixer llocs d'estroncament (principis/finals d'exons) és la cadena de Markov d'ordre 0 o 1 específica de posició (Michael R Brent & Roderic Guigó, 2004). La cadena de Markov d'ordre 0 és un model en el que cada posició relativa a lloc d'estroncament (-1,+1, +2...) té una probabilitat assignada a cada base (A,C,G,T) de forma independent a la resta. El resultat és una matriu de pesos (PWM; de l'anglès *position weight matrix*) construïda a partir d'un conjunt de seqüències funcionals, que puntua la probabilitat de que la seqüència sigui un lloc actiu (veure figura 4.4). Aixi, per exemple, Geneid reconeix primer els codons d'inici i de final i les senyals d'estroncament usant PWA (*position weight array*) , i en una segona fase els exons sencers són definits i avaluats. Finalment la estructura final del gens es construeix i es re-avalua (Parra, Blanco, & Roderic Guigo, 2000).

Les cadenes de Markov de primer ordre, són un model que condiciona la probabilitat de cada posició i base, en funció de la base en la posició anterior. Genscan incorpora un model una mica més complex que incorpora dependencies de posicions no adjacents (Burge & Karlin, 1997). Altres mètodes matemàtics que són usats per capturar les senyals en la seqüència que determinen la funcionalitat d'una regió són el IDQD (*increment of diversity quadratic discriminant analysis*), els *support vectors machines* (SVM), els models d'entropia màxima o el les xarxes bayesianes (Michael R Brent & Roderic Guigó, 2004)

4.1.2.1.2 Mètodes basats en l'anàlisi comparatius

Donat el número elevat de genomes de diferents especies que han estat seqüencials, un estratègia molt interessant per la predicció de gens són els mètodes d'anàlisi comparatiu entre genomes. Aquests es basen en el principi que la selecció natural causa en els gens i altres elements funcionals d'acumular menys mutacions que en la resta del genoma, ja que les mutació en un llocs funcionals són més properes a tenir un impacte negatiu en l'organisme. En conclusió, els predictors assumeixen que els elements funcionals del genoma (p.ex. exons) estaran més conservats en altres espècies que els no funcionals

(p.ex. introns). Després del seqüenciament del genoma del ratolí (Sequencing Consortium, 2002) el genoma humà disposava d'un genoma suficientment proper al qual poder-se comparar. És en aquest moment que aquesta estratègia va ser adoptada per programes com Slam (Alexandersson, Cawley, & Pachter, 2003), Sgp-1 (Wiehe, Gebauer-Jung, Mitchell-Olds, & R Guigó, 2001) o Twinscan/Nscan (I Korf, P Flicek, D Duan, & M R Brent, 2001b), contribuint notablement a l'anotació del genoma humà.

Twinscan per exemple, incorpora el model de probabilitat *ab initio* per la predicció de la estructura dels gens de Genscan amb la informació de conservació d'una regió entre dos genomes. El predictor entrena de forma independent la conservació en les regions d'exons, introns, llocs d'estroncament i UTR, posant de relleu les diferencies entre ells (I. Korf, P. Flicek, D. Duan, & M. R. Brent, 2001). Aquest programa demostra la millora de la incorporació d'un anàlisi comparatiu en la sensitivitat i la especificitat (veure secció Avaluació dels mètodes predictors) del predictor.

4.1.2.1.3 Mètodes basats en homologia

Una informació que dona un gran avantatge en la predicció dels gens en un genoma, és tenir el coneixement dels RNAms que transcriuen o les proteïnes que tradueixen. Gracies a l'avenç en tècniques de seqüenciació massiva que son comercialitzades per més d'una companyia *(Chi, 2008)* avui en dia és una realitat poder disposar del mapa complert de les seqüències de ARNms expressades en un teixit.

Els problemes principals per identificar quina regió genòmica transcriu un ARNm o codifica una proteïna són:

- Una proteïna potser codificada a partir de múltiples seqüències d'ADN .
- En eucariotes tant la proteïna com l'ARNm madur no tenen les regions intròniques, que poden ser molt llargues, introduint soroll a l'anotació del gen.
- Donada la mida dels genomes (3 gigues en humans) i el número de seqüències

(mes de 10,000 proteïnes en humans) el cost computacional pot ser molt gran.

- Donat que l'origen del la proteïna/ARNm pot ser d'un individu diferent al qui s'ha seqüenciat el genoma, poden existir petites variacions degudes a SNPs, que introdueixin soroll a l'anotació.

Aquests quatre punts no suposen cap problema però per algoritmes que s'han especialitzat en aquesta tasca com es el cas de Blat (W. James Kent, 2002). Blat és una eina d'alineament entre seqüències similar a la popular eina Blast (Jian Ye, McGinnis, & Madden, 2006) que busca regions de semblança local entre dues seqüències. Però així com, Blast està orientada a compara dues seqüències semblants (proteïna vs proteïna), Blat busca seqüències en genomes sencers (proteïna, genoma). Blat és capaç de reconèixer la seqüència entrant com ADN, ARN o proteïna, essent capaç d'interpretar que la seqüència del ARN madur es trobarà en el genoma intercalada per introns. Les seves prediccions són en general precises i ràpides.

4.1.2.2 Avaluació dels mètodes predictors

La avaluació dels resultats d'un programa predictor és una eina fonamental, no només per millorar les estratègies existents, sinó per validar la hipòtesis de treball en la que es basa un programa específic (Burset & R Guigó, 1996). Una avaluació acurada requereix:

- D'un conjunt d'elements coneguts i validats de forma externa, que no hagin estat utilitzats en l'entrenament del programa predictor, en aquest cas un grup de gens, que coneixem com a conjunt test.

- Una mètrica de valoració, ja que després de la predicció del conjunt test, existirà en la majoria dels casos una disconformitat entre aquest i el conjunt de candidats predits pel programa. Aquesta disconformitat entre els valors reals (grup test) i els valors predits es pot comptabilitzar a partir dels paràmetres següents (veure Figura 4.5):

- Positius Verdaders (TP; *true positive*): prediccions que coincideixen amb la realitat

- Positius Falsos (FP; *false positive*): prediccions que no coincideixen amb la realitat

- Negatius Falsos (FN; *false negative*): elements reals sense predicció

- Negatius Verdaders (TN; *true negative*): tota la resta

- Si s'avaluen de forma conjunta diferents predictors de forma simultània, tots ells han de respectar el mateix protocol, i per tant, cap d'ells pot haver usat el conjunt test en un entrenament i cal que la definició de TP, FP, FN i TN sigui exactament la mateixa.

La manera com es decideixi definir els 4 paràmetres d'una predicció (TP, FP, FN i TN) definiran el comportament del programa predictor. Si això s'ha fet de forma correcta TP + FP = total de prediccions, TP + FN = total d'elements reals en el conjunt test i el total de candidats és igual a Ntot = TP+ TN+ FP+ FN.

En la predicció de gens aquests termes poden interpretar-se de maneres diferents (Burset & R Guigó, 1996) una d'elles és a partir de nucleòtids. En aquest cas Ntot és la longitud del genoma (o de la regió del genoma que s'avaluï) i TP+FP és el número total de nucleòtids en exons predits. Per altra banda, TP+FN és el número total de nucleòtids en el conjunt de test de gens (veure Figura 4.5)

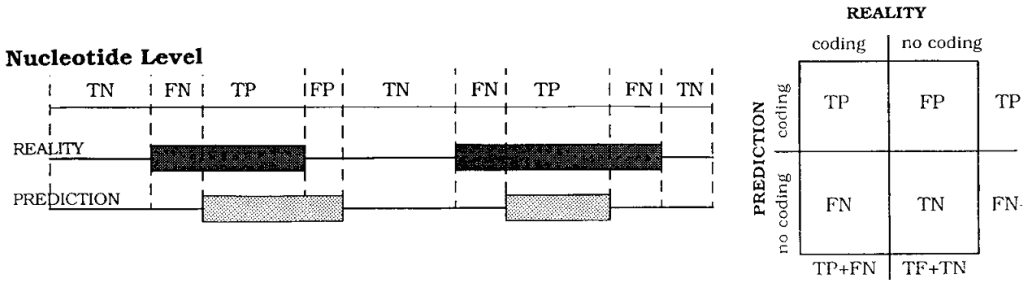

Figura 4.5 Determinació dels paràmetres de predicció (TP, FP, TN, FN) a partir de nucleòtids

segons Burset et al. (Burset & R Guigó, 1996).

Un cop calculats els paràmetres, els elements que millor descriuen un predictor són la sensibilitat, la especificitat i la proporció de prediccions correctes. La sensibilitat (SENS) d'un predictor avalua el percentatge de nucleòtids en gens reals que són correctament capturats (veure equació 4.1). La especificitat (SPEC) mostra el percentatge de nucleòtids sense predicció de gens encertats (veure equació 4.2). Finalment la proporció de prediccions correctes (PPV; *proportion of correct predictions*), coneguda també com segon coeficient d'especificitat, avalua el percentatge de prediccions efectuades que són finalment reals (veure equació 4.3).

$$SENS=\frac{TP}{TP+FN} \qquad (4.1)$$

$$SPEC=\frac{TN}{TN+FP} \qquad (4.2)$$

$$PPV=\frac{TP}{TP+FP} \qquad (4.3)$$

Entre aquest tres descriptors són la SENS i el PPV els que millor avaluen un mètode i la manera de combinar-los per obtenir una sola puntuació (Burset & R Guigó, 1996) és el coeficient de correlació de Mathews (CC, veure equació 4.4).

$$CC=\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TF+FN)(TN+FP)(TN+FN)}} \qquad (4.4)$$

Apart de les mesures estàndard descrites, existeixen altres indicadors de la precisió d'un programa predictor, com per exemple el promig d'errors en les prediccions (AE, en anglès *average mismatch of predictions*; veure equació 4.5), el coeficient associat de Yule (Q; veure equació 4.6) o les distàncies generalitzades als predictors ideals (GDIP1, GDIP2 i GDIP3; en anglès *generalized distances from ideal predictors*; veure equacions 4.7, 4.8 i 4.9 respectivament). Finalment un mètode per combinar aquestes i altres mesures és el promig de les puntuacions de les mesures (ASM; en anglès

averaged score measure; veure equació 4.10) que integra de manera independent cada un dels descriptors per aconseguir una estimació global del poder predictiu de l'algoritme, respecte a altres solucions.

$$K2 = \frac{TP + TN}{FN + FP + \frac{1}{N_{tot}}} \tag{4.5}$$

$$Q = \frac{(TP \times TN) - (FP \times FN)}{(TP \times TN) + (TP \times TN)} \tag{4.6}$$

$$GDIP1 = \frac{\sqrt{FP^2 + FN^2}}{TP + TN + \frac{1}{N_{tot}}} \tag{4.7}$$

$$GDIP2 = \frac{\sqrt{FP^2 + FN^2}}{TP + \frac{1}{N_{tot}}} \tag{4.8}$$

$$GDIP3 = \frac{\sqrt{FP^2 + FN^2}}{TN + \frac{1}{N_{tot}}} \tag{4.9}$$

$$ASM_i = \frac{1}{z} \sum_{j=1}^{z} P_j^i \tag{4.10}$$

On z es el nombre de mesures diferents usades (SENS, SPEC, PPV, Q, …) que ordenen de forma independent el p programes avaluats. Definim aquí per el mètode i usat el vector de programes ordenats com $r_i = [P^i_1,..., P^i_z]$.

4.1.2.3 Predicció de promotors

La polimerasa II (Pol-II) és l'element principal que determina la transcripció de la majoria dels gens (veure capítol 2). També sembla força evident que el complex Pol-II no és capaç per si mateix d'iniciar la transcripció, sinó que necessita un nombre

addicional de proteïnes conegudes com factors de transcripció (TF) que solen unir-se a la regió del promotor mes proximal (250 bps) al inici de transcripció. Els primers anàlisis de senyals comuns en el promotor proximal dels gens coneguts en eucariotes van revelar la presència d'alguns motius estadísticament sobre-representats als quals es va associar al posicionament de complex iniciador de la transcripció. La caixa TATA (Bucher, 1990) és la més coneguda d'aquestes. La seqüència consens de la caixa TATA és TATAAAA, però la variabilitat d'aquest motiu en els diferents gens és molt gran (Smale & Kadonaga, 2003). La posició d'aquest motiu es sol trobar a una distància de 25-30 parells de bases abans del inici de transcripció, però aquest valor també pot variar de forma significativa, especialment entre espècies (K Struhl, 1989). Tradicionalment es suposava que la caixa TATA es trobava en un 30% dels gens, però no està present en gens de manteniment (gens essencials per la viabilitat de la cèl·lula; en anglès *housekeeping*), tampoc en factors de creixement ni oncogens (Solovyev & Shahmuradov, 2003b). De fer, experiments de detecció massiva d'inicis de transcripció de forma exhaustiva en mamífers pel mètode *cap analysis of gene expression* (CAGE) (Carninci et al., 2006; Kim et al., 2005) han revelat que el nombre de TSS relacionats amb una caixa TATA a mamífers és clarament inferior al que els experiments amb especies inferiors suggerien.

Un segon element tradicionalment conegut en promotors és l'Element Iniciador (Initiator Element en anglès; Inr), localitzat al voltant del TSS. La seqüència consens establerta per l'Inr en diferents organismes és A/GGA/TC/TG/A/C i es localitza en posicions +28 to +32 respecte al TSS (J. E. F. Butler & Kadonaga, 2002). No està clara la prevalença de l'Inr en mamífers, però sembla abundant en la mosca on es podria trobar al voltant d'un 40% dels gens (Smale & Kadonaga, 2003).

Es coneix que les illes CpG (veure capítol 2) estan presents en aproximadament la meitat de gens humans, amb especial representativitat entre els *housekeeping* (Bajic & Seah, 2003; Gardiner-Garden & Frommer, 1987; Pedersen, Baldi, Chauvin, & Brunak, 1999; M. Q. Zhang, 2002). Les illes CpG no tenen un motiu de consens, però són detectades com llargs (500~2000 parells de bases) segments d'ADN amb un alt contigut de C+G i una població alta del dinucleòtid CpG (la definició matemàtica d'una illa CpG majoritariament acceptada (F. Larsen, Gundersen, R. Lopez, & Prydz, 1992)

és: per a una finestra n >200 cal que es compleixin les condicions i) 0.5 < (#C + #G)/n, ii) 0.6 < Obs/Exp; on Obs = #CG i Exp = (# C *#G) / n). El promotors associats a illes CpG solen tenir múltiples llocs d'inici de transcripció que es distribueixen en centenars de bases de distància (Davuluri, Grosse, & M. Q. Zhang, 2001) i que rarament tenen una caixa TATA o elements Inr (Y. Suzuki et al., 2001).

Tot i els impressionants avanços en les tècniques d'alt rendiment per la detecció de TSS (Carninci et al., 2006; Kim et al., 2005) la determinació de promotors és encara una tasca difícil, el que fa necessari recorre a mètodes de predicció. En principi, aquesta predicció seria fàcil si el gen al que correspon el promotor està perfectament anotat i aquest regió està localitzada en les bases immediatament anteriors a l'extrem terminal 5'. Malauradament, l'anotació està subjecta encara a grans incerteses i recents anàlisis genòmics han demostrat que el concepte simplificat d'un gen amb un sol TSS localitzat corrent amunt de 5' (*5'-upstream*) no és vàlid. Els TSS poden estar repartits per tot arreu, fins i tot en la regió transcrita d'un gen (incloent introns, exons i regions terminals no transcrites). El model es complica si tenim en compte de que un sol promotor pot contenir diversos TSS i que sovint dues regions promotores diferents es solapen (Bajic et al., 2006; Carninci et al., 2005, 2006). En altres paraules, la predicció teòrica de promotors és encara un dels grans reptes de la bioinformàtica.

Nscan, una versió modificada de *Twinscan* (Gross & M. R. Brent, 2006), és el paradigma d'un mètode de predicció de promotors basat en la detecció de la estructura del gen. El programa usa cadenes de Markov ocultes (*Hidden Markov Model*; HHM) per localitzar les senyals que es troben en els gens i localitza el terminal 5' del gen regions conservades a través d'altres espècies, per tenir així una idea del posicionament del promotor (veure figura 4.6). Una alternativa als mètodes com Nscan basats en la conservació de la estructura gènica són els que tracten de recuperar la posició del promotor a partir de senyals subtils a nivell de seqüència que poden ser associades al funcionament del promotor. D'aquesta manera, molts mètodes estan entrenats per detectar motius de seqüència; com per exemple, la caixa TATA, l'Inr, una illa CpG o un població anormalment alta de llocs d'unió de TFs (Bajic & Seah, 2003; Bajic et al., 2003; Davuluri et al., 2001; T. A. Down & T. J. Hubbard, 2002; Knudsen, 1999; Ohler et al., 2002; Ponger & Mouchiroud, 2002; M. G. Reese, 2001; Solovyev & Salamov,

1997). Donada la gran representativitat d'illes CpG en promotors de mamífers, molts mètodes han usat l'estratègia d'entrenar-se individualment pels dos tipus de regions reguladores: CpG+ o CpG- (Bajic & Seah, 2003; Bajic et al., 2003; Davuluri et al., 2001). Molts d'aquests mètodes usen regles de composició a nivell de tri-nucleòtids (o pentàmers, hexàmers, ...) o sofisticades versions de HMM entrenades contra promotors coneguts. Finalment, alguns mètodes com *Firstef* (Davuluri et al., 2001), *Dragon promoter finder* (Bajic & Seah, 2003) o *PromoH* (Solovyev & Shahmuradov, 2003b) combinen tècniques de predicció d'estructura gènica amb models de reconeixement d'elements de seqüència.

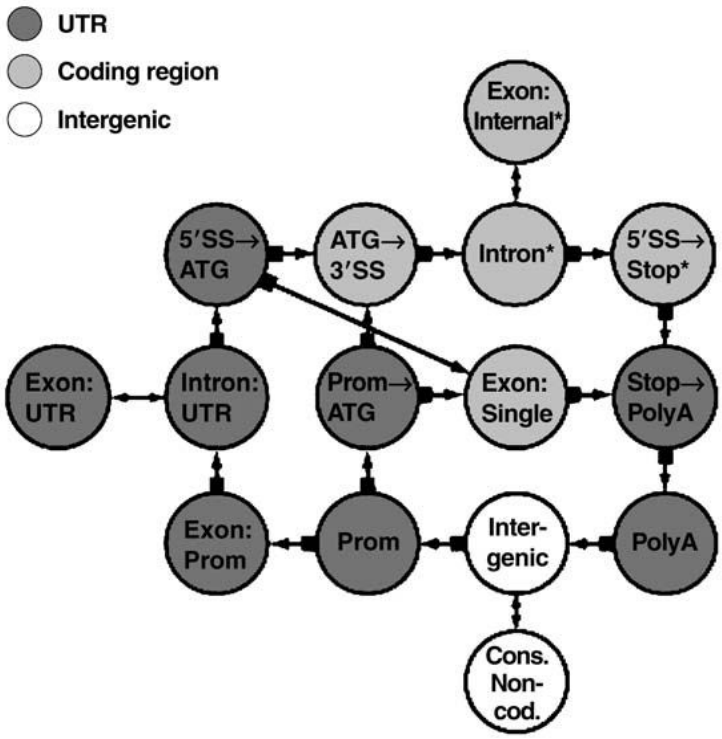

Figura 4.6 El diagrama d'estats de Nscan. El programa determina la predicció d'un promotor a partir de la predicció de gen (exons i introns). El programa es complementa finalment amb l'estudi de la conservació de les regions entre especies.

El caràcter difús de les senyals de seqüència en els promotors indiquen que altres factors apart del reconeixement de seqüència per ponts d'hidròxens podrien regular el reconeixement de fragments d'ADN per part de les proteïnes. Tal com ja ha estat

suggerit (Kanhere & Bansal, 2005; Pedersen et al., 2000) un d'aquest factors addicionals és el de les propietats físiques de la molècula d'ADN, que controla el grau d'accessibilitat de les seqüències objectiu per part dels factors de transcripció. El fet que l'ADN en regions promotores mostri unes propietats físiques que a la resta del genoma, especialment a prop del TSS, havia esta provat de forma clara en procariotes i amb menor mesura en alguns eucariotes (Florquin et al., 2005; Kanhere & Bansal, 2005; Ohler, Nierman, Liao, & G. M. Rubin, 2001; Pedersen, Baldi, Chauvin, & Brunak, 1998, 1999; Pedersen et al., 2000). Un dels objectius de la tesis ha estat aprofundir en aquest punt i analitzar la possibilitat de determinar regions promotores reconeixent les seves propietats físiques inusuals.

4.1.2.3.1 Avaluació dels predictors de promotors

L'avaluació del comportament dels programes predictors segueix la mateixa lògica que la dels programes predictors de gens. La diferència principal radica en la definició dels paràmetres TP, FP, TN, FN. Recentment el grup de treball Egasp (M. G. Reese & R. Guigo, 2006) va assentar un protocol per establir de forma fiable una anàlisis comparatiu dels predictors en el genoma Humà que s'ha convertit en un estàndard. Egasp proposa usar les prediccions curades de TSS fetes pel grup Havana ("The HAVANA Team") en la regió Encode (T. E. P. Consortium, 2004). Seguint aquestes indicacions, els TP són els TSS d'Havana amb una predicció en el mateix sentit dins d'una distància màxima de D nucleòtids (on es proposa D=1000 o D=250). Si la anotació del TSS no té cap predicció propera, es computada com FN. Cada una de les prediccions que estigui en la regió transcrita (i en el mateix sentit) d'algun dels gens d'Havana (sense que estigui a prop del TSS; de la posició Gen_inici+D a Gen_final) compta com un FP. Els TN són la suma de posicions dins la regió transcrita del gen (Gen_inici+D a Gene_final) que no tenen cap predicció (M. G. Reese & R. Guigo, 2006).

4.1.2.4 Predicció de llocs d'unió de factors de transcripció

La expressió dels gens està regulada, entre d'altres elements, per factors de transcripció (veure capítol Biologia Molecular). La majoria de TF tenen la habilitat de reconèixer i enganxar-se a petits (6-15 nucleòtids) segments d'ADN per activar, potenciar o

reprimir la transcripció d'un gen. Malauradament els motius d'unió per a un TF específic pot ser altament variable, el que dificulta la seva detecció al genoma.

La tècnica de Chip-on-chip (veure figura 4.7) és la més emprada per detectar llocs d'unió a factors de transcripció. En un experiment de Chip-on-Chip primer es fragmenta la cromatina i s'exposa a un anticòs específic al FT d'interès. Aquests complexos es precipiten i posteriorment s'aïlla l'ADN que es trobava enganxat al TF. Usant la tècnica de *microarrays* d'ADN és posa de relleu quines són les seqüències genòmiques on hi havia interacció física amb el TF (veure figura 4.7). Aquesta tècnica millora sensiblement amb l'ús de plataformes d'última generació com els *tilling microarrays*, amb els quals s'han detectat de forma precisa TFBS en genomes de la llevat, la mosca i cèl·lules de mamífers (Lander, 1999; O'Geen et al., 2007). La predicció i/o validació experimental de llocs d'unió a un TF és però, una tasca molt costosa i no sempre 100% efectiva: no només per la necessitat de tenir anticossos específics contra els diferent factors de transcripció, sinó per que el no obtenir cap senyal de que una posició sigui lloc d'unió per un TF conegut no demostra que no sigui actiu en unes altres condicions cel·lulars.

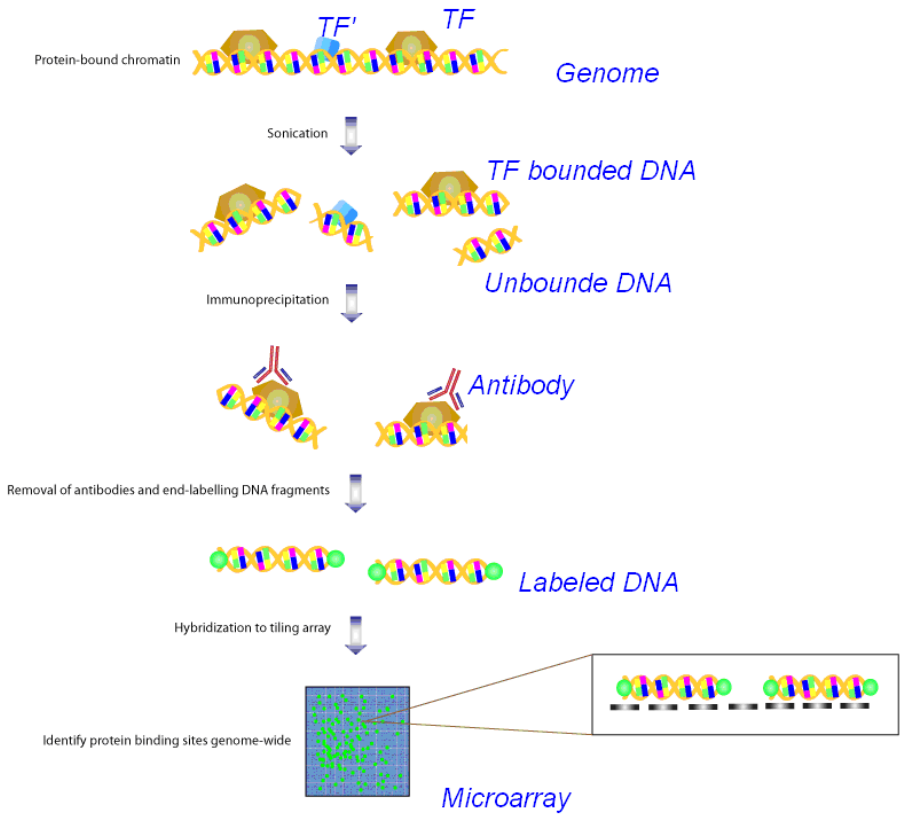

Figura 4.7 Experiment Chip-on-chip per localitzar els llocs d'unió d'un TF d'interès a nivell genòmic. Cal realitzar l'experiment cada cop per un TF independent cosa que el fa un sistema molt costos.

La predicció de llocs funcionals d'unió a TF computacionalment és un problema que pot ser resolt de manera semblant a la predicció de gens. Existeixen bases de dades que recullen la col·lecció (veure secció Bases de dades bioinformàtiques) de motius d'unió validats per a un TF específic (Albin Sandelin, Alkema, Engström, Wasserman, & Boris Lenhard, 2004; Wingender et al., 2001). Un popular sistema per captura els TFBS és doncs l'ús de PWM establertes a partir d'aquests motius. És el cas de Promo (Messeguer et al., 2002) una eina bioinformàtica que pre-calcula les matrius de pesos a partir dels llocs d'unió coneguts en cada espècie, localitza objectius d'unió en un segment d'ADN genòmic i assigna a cada predicció un valor estadístic per determinar la versemblança del candidat. El resultat però, és que donada la flexibilitat en la

seqüència consens dels TF, aquesta estratègia pot donar un gran nombre de falsos positius. Un sistema per corregir aquest efecte, igualment usat en la predicció de gens, es l'ús d'un anàlisi comparatiu entre espècies per millorar la llista de candidats de TFBS. És el cas de la base de dades ABS (Blanco, Farré, M. Mar Albà, Messeguer, & Roderic Guigó, 2006) que fa ús de la conservació entre espècies per filtrar un conjunt de TFBS extrets de la literatura, assumint que els llocs funcionals estan més ben conservats en espècies properes

Les estratègies anteriors però, només són aplicables quan del TF se'n coneix el lloc d'unió, però és força comú el cas de que aquest no sigui conegut. En aquests casos la estratègia es pot basar en l'assumpció de que els llocs funcionals, a diferència dels no funcionals, es trobarà en forma de motius similars repartits en múltiples promotors. MEME (T. L. Bailey, Nadya Williams, Misleh, & Wilfred W. Li, 2006) és una de les eines més populars en la cerca de noves senyals de seqüència a partir d'un grup de fragments d'ADN. MEME és usada per la localització de TFBS o dominis de proteïna. El seu funcionament es basa en la cerca de motius repetits (amb una similitud molt alta) que es donen en l'ADN (o proteïna) que l'usuari ha de facilitar.

4.1.2.5 Predicció seqüències posicionadores de nucleosomes

El posicionament dels nucleosomes en el genoma és un element dinàmic que intervé directament en la regulació dels gens (veure capítol Biologia molecular). El principi d'interacció entre ADN i nucleosomes és idèntic al dels factors de transcripció. Les estratègies que funcionen però en la predicció del TFBS no són aplicables però al nucleosomes perquè:

- La unió de les histones a l'ADN no té cap senyal aparent. Ni tant sols entre espècies es troben senyals específiques.

- És difícil buscar elements repetits que es donen en seqüències nucleosomals quan els nucleosomes cobreixen els organismes superiors per complet.

Els primers intents per desenvolupar un programa predictor de regions ocupades per nucleosomes, tenen ja més d'una dècada (Brukner et al., 1995; Satchwell, Drew, &

Travers, 1986; Sivolob & Khrapunov, 1995a). La digestió de l'ADN per mitjà de l'enzim DNasa I i la nucleasa micrococal, és una via usada per aïllar seqüències unides a nucleosomes, ja que aquestes, protegides pel complex, tarden mes a degradar-se. Un cop aïllades i seqüenciades es poden derivar els paràmetres basats en una taula de freqüències de dinucleòtids (Sivolob & Khrapunov, 1995b) o trinucleòtids (Brukner et al., 1995). Tot i que, el més important és conèixer la posició ocupada pel nucleosoma (posició translacional) , també ho és conèixer la orientació i l'encaix d'aquest en aquesta seqüència, coneguda com posició rotacional. Existeix una taula de paràmetres de dinucleòtids derivats a partir de 177 seqüències de pollastre ocupades per nucleosomes que permet predir la orientació del nucleosoma en una seqüència específica (Satchwell et al., 1986).

Recentment s'han desenvolupat predictors més sofisticat com Recon (V G Levitsky, Podkolodnaya, N A Kolchanov, & Podkolodny, 2001). Recon s'entrena a partir de dos grups de seqüències: seqüències formadores de nucleosomes validades experimentalment (~150 seqüències) (Ioshikhes & E N Trifonov, 1993) contra seqüències aleatòries amb certes normes de composició genòmica. El programa busca diferències significatives en seccions de les seqüències que distingeixin de forma òptima els dos grups a través del càlcul de la distància de Mahalanobis (Mahalanobis, 1936). La definició de les seccions en les seqüències es determina a partir d'un entrenament (V G Levitsky et al., 2001) realitzat amb un algoritme de Monte Carlo (veure seccions següents).

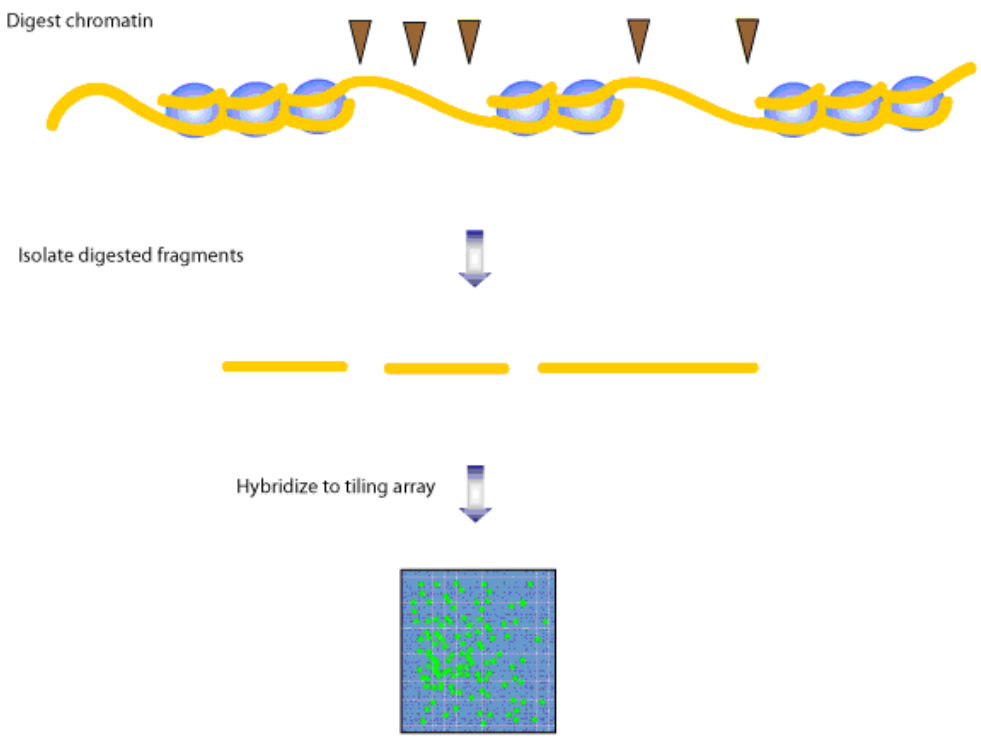

Figura 4.8 Esquema de la tècnica experimental per la localització de seqüències ocupades en el genoma *in vivo*. Es coneix que la degradació de l'ADN és menor en segments ocupats per nucleosomes, pel que calculant be els temps és possible aïllar i localitzar les seqüències on es posicionen.

En el cas dels nucleosomes les tècniques de Chip-on-chip són també aplicables tot i que també està molt extensa la tècnica de digerir l'ADN (veure figura 4.8) per posteriorment analitzar els resultats amb un microarray (Guo-Cheng Yuan et al., 2005). En aquest cas, s'ha aconseguit identificar la posició 2,278 nucleosomes en quasi 500 Kbases que inclouen principalment el cromosoma III de la *Saccharomyces cerevisiae*. L'estudi posa de relleu la gran sinergia entre els inicis de transcripció i el posicionament del nucleosoma. La part que no queda resolta però, es que tot i que la majoria de nucleosomes trobats estan ben posicionats (Guo-Cheng Yuan et al., 2005), no sabem quins d'ells ho fan perquè tenen una forta preferència per la seqüència o perquè és l'únic lloc que els queda lliure. Recentment, l'ús de tècniques de digestió

inespecífica de segments inter-nucleosomals s'ha barrejat amb les seqüenciació massiva per obtenir el posicionament massiu de nucleosomes en diferents espècies (Schones et al., 2008).

Existeix la hipòtesis de que algunes seqüències tenen la propietat de ser fortament posicionadores de nucleosomes i a partir d'elles la resta de nucleosomes veïns s'ajusten de forma ordenada (Segal et al., 2006). Així doncs, per usar mètodes computacionals eficients no és suficient la definició dels paràmetres predictors de seqüència afí a histones, sinó que cal que el programa tingui en compte el perfil global dels potencials al llarg d'un segment d'ordre superior. Els resultats del mètode que integra aquesta informació millora notablement els resultats realitzant prediccions molt ben ajustades a la realitat tal com s'aprecia en la figura 4.9.

En els últims anys la millora de qualitat en les dades experimentals en llebat (Guo-Cheng Yuan et al., 2005), mosca (Mito, J. G. Henikoff, & S. Henikoff, 2005, 2007) o humà (Dennis et al., 2007; Ozsolak, Song, X Shirley Liu, & Fisher, 2007) han millorat notablement el poder predictius dels algoritmes *in silico*. Aquesta nova generació de predictors basats en SVM o HMM (Gupta et al., 2008; Miele, Vaillant, d'Aubenton-Carafa, Thermes, & Grange, 2008; Peckham et al., 2007) han demostrat la relació entre les seqüències posicionadores de nucleosomes i paràmetres estructurals de l'ADN com la curvatura o la flexibilitat intrínseca de l'ADN. Aquest fet posa de manifest la necessitat de disposar d'eines bioinformàtiques que reconeguin característiques estructurals a partir de la seqüència de dinucleòtids.

Per últim no podem oblidar que la cromatina és dinàmica i que possiblement el grau de compactació en nucleosomes i el propi posicionament dels mateixos pot canviar molt depenent de l'estat d'activitat dels gens (Schones et al., 2008).

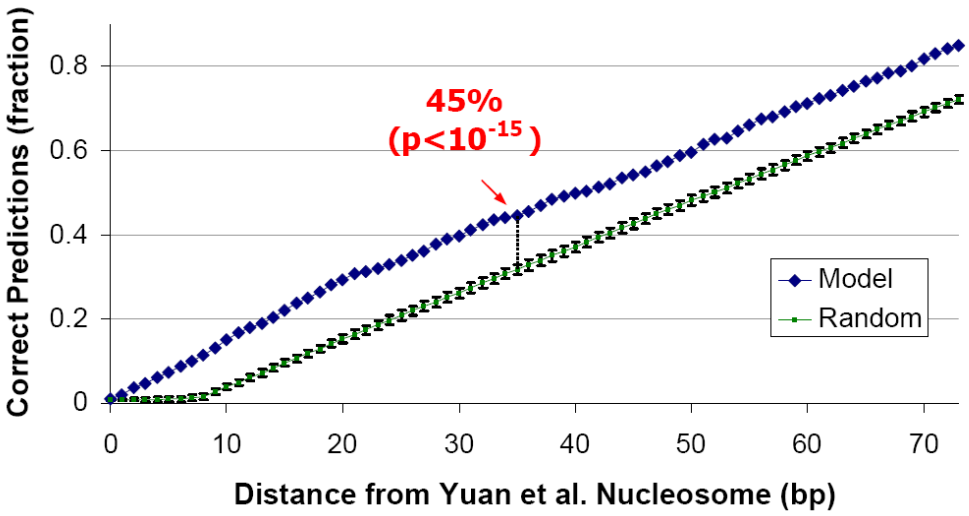

Figura 4.9 Resultats del programa predictor de posicions dels nucleosomes desenvolupat pel grup de Widom i col·laboradors (Segal et al., 2006) i validat amb les dades experimentals de *Saccaromyces* (Guo-Cheng Yuan et al., 2005). En la gràfica mostra número de prediccions correctes (eix-Y) assumint un error en nucleoòtids (eix-X) de les posicions real. La gràfica compara els resultats del programa contra un model d'atzar, a partir del qual es pot calcular la significança estadística del poder predictiu (en vermell).

4.1.3 Mètodes de predicció de la dinàmica de l'ADN

L'estudi de la dinàmica de la molècula d'ADN es pot fer a molts nivells ja sigui macroscòpic (ideal-elàstics), mesoscòpic (entremitjos) o microscòpic (atòmics). Per interès d'aquesta tesis solament s'introduiran alguns dels mètodes microscòpics. Dins d'aquesta categoria existeixen dos grups: els basats en variables col·lectives i els basats en el model atòmic. Els primers es basen en el coneixement previ de diversos paràmetres relacionats amb la estructura dels àcids nucleics, els quals permeten reduir els graus de la llibertat configuracional no explicits. Un exemple d'aquests seria els models helicoïdals (veure capítol Estructura de l'ADN), on la estructura es representa per una sèrie de paràmetres rotacionals i translacionals que fan referència a nivell de un parell de bases o a bases aïllades. Els models microscòpics atòmics, no obstant,

utilitzen una descripció detallada de tots els graus de llibertat del sistema, usualment en l'espai cartesià.

4.1.3.1 Dinàmica molecular

La Dinàmica Molecular (MD) és una tècnica determinista, que permet predir l'estat següent del sistema a partir de l'estat actual, definit en funció de les seves posicions i moments. Els seus fonaments es basen en les lleis de la mecànica clàssica, integrades en equacions de moviment, amb la finalitat de generar successives configuracions i permetent així obtenir una seqüència temporal de moviments de la evolució del sistema. Les configuracions s'obtenen solucionant la equació diferencial continguda en la segona llei de Newton (equació 4.11 i figura 4.10) que descriuen el moviment d'una partícula i de massa m_i al llarg de les coordenades x_i, essent F_i la força aplicada (obtinguda per diferenciació a l'espai de la energia potencial clàssica del sistema) i t la variable del temps.

$$\frac{d^2x_i}{dt^2} = \frac{F_i}{m_i} \qquad (4.11)$$

La força es determinada per diferenciació analítica de la energia potencial (determinada típicament a nivell clàssic) respecte al moviments del àtoms (equació 4.12).

$$F_i = -\frac{dE}{dx_i} \qquad (4.12)$$

Un com es coneixen les accelarió per integració simple es deriven les velocitats i per integración d'aquestes les noves possicions. La complexitat de la superfície potencial requereix que les equacions del moviment siguin integrades numèricament en petits passos consecutius (aproximadament cada femtosegon). En cada pas es calculen les forces dels àtoms i posteriorment es combinen les posicions actuals per generar les noves posicions. Els àtoms són desplaçats a les seves noves posicions i es torna a començar el cicle (veure figura 4.10).

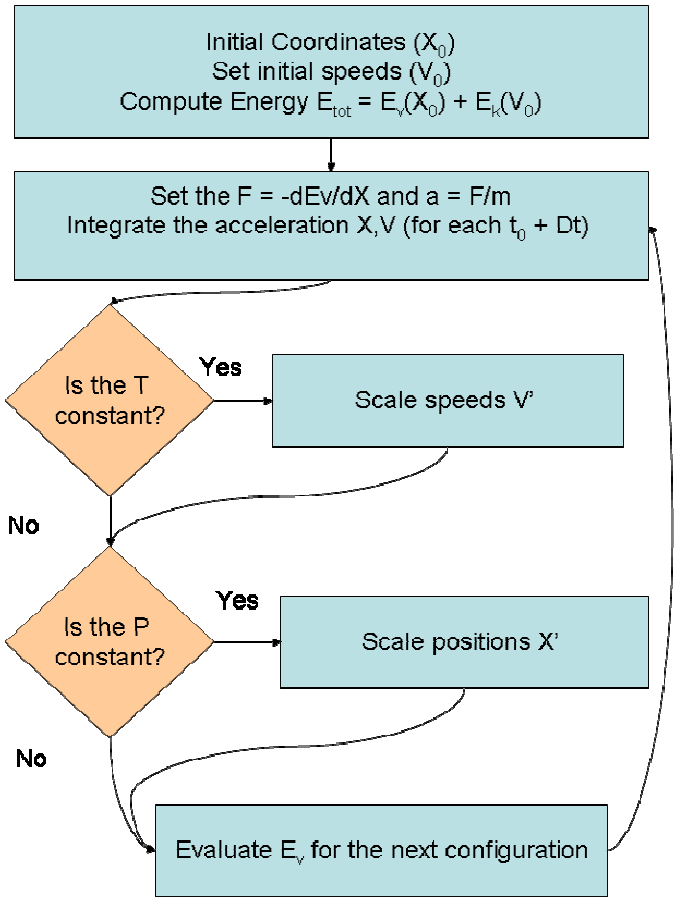

Figura 4.10 Algoritme bàsic de la dinàmica molecular.

Tal com s'ha introduït, l'última finalitat de les simulacions d'àcids nucleics és la de reproduir les propietats macroscòpiques a partir de models atòmics. Així doncs, per reproduir les propietats fisiològiques de l'ADN, a més de mantenir la temperatura i pressió dins d'uns rangs (veure figura 4.10) cal que integri també explícitament l'efecte del solvent (aigua en la majoria dels casos). La introducció de les molècules del solvent al sistema incrementa molt els temps de càlcul de les MD, limitant a la pràctica la mida del sistema a simular (típicament a unes desenes de milers d'àtoms), i la llargada de la simulació (típicament unes desenes de nanosegons),

Preparació d'una Dinàmica Molecular

El punt inicial de qualsevol simulació és sempre una configuració inicial (veure figura 4.10). Aquesta estructura pot provenir de dades experimentals; com per exemple: raig-X, ressonància magnètica nuclear o d'un model teòric. Així mateix es requereix d'un conjunt inicial de velocitats, que s'assignaran a l'atzar seguint una distribució del tipus Maxwell-Boltzmann. Després de la primera integració del equació de Newton es passa a un procés d'equilibrat, el propòsit del qual es eliminar distorsions en la trajectòria degudes a imprecisions a l'estructura original o amb el conjunt de velocitats aleatòries que es fan servir per la primera integració de les equacions de Newton. Durant aquest període es solen controlar amb detall els paràmetres del sistema, fins que s'hagin estabilitzat. La duració d'aquest període es variable i depèn en gran mesura del sistema d'estudi. En el cas dels àcids nucleics es pot arribar a una situació d'equilibri en aproximadament 1 nanosegon (Manuel Rueda, 2005).

Un cop superat l'equilibrat s'entra en la fase de producció de la dinàmica. La duració de la mateixa serà sempre un compromís entre els recursos computacionals, espai de disc i la capacitat de racionalitzar els resultats (Manuel Rueda, 2005), l'estat de l'art actual implica simulacions al voltant dels 50-100 ns, però ja ha estat publicada la primera simulació en el rang del microsegon per l'ADN (Pérez, F Javier Luque, & Modesto Orozco, 2007). La evolució de la dinàmica de la estructura d'ADN quedarà recollida en forma de coordenades cartesianes, que posteriorment es poden processar (veure secció Recollida de dades: dinàmica essencial)

4.1.3.1.1 Camps de força

Els camps de força es poden definir com la dependència empírica de la energia potencial respecte a la geometria dels nuclis. Són potencials efectius ajustats per reproduir càlculs quàntics acurats i dades experimentals en sistemes de mida gran sempre que no hi hagi alteració de enllaços covalents ni canvis dràstics en la distribució electrònica. Un camp de força està composat per dos components diferents.

- Un conjunt d'equacions, anomenades funcions potencials, usades per determinar la energia potencial a partir de la estructura.

- Els paràmetres empírics usats per les equacions.

És possible utilitzar diferents paràmetres amb el mateix conjunt d'equacions, encara que és aconsellable comprovar la correspondència entre els mateixos. Existeixen diversos camps de força de (M Orozco et al., 2004), essent les famílies de camps de força AMBER (Weiner, Kollman, D. T. Nguyen, & Case, 1986) i CHARMM (Brooks et al., 1983) els més utilitzats en el camp dels àcids nucleics.

4.1.3.2 Simulacions amb Monte Carlo

No existeix un sol mètode de Monte Carlo (MC), sinó que aquest terme descriu una llarga llista d'algoritmes amb variants molt utilitzades en problemes molt diversos. Es tendeix a recórrer als mètodes de Monte Carlo quan un problema no és assumible pel seu cost o en cas que no sigui computable el resultat exacte per un algoritme determinista. És el cas de l'exploració de les dinàmiques de segments genòmics d'ADN, impossibles de determinar per MD atomística amb la tecnologia actual. Qualsevol de les variants d'un Monte Carlo compleix els següent patró (Metropolis & Ulam, 1949):

1. Es defineix un domini de possibles entrades al sistema

2. Es generen entrades (del domini definit) de forma aleatòria i es realitza un càlcul determinista en elles.

3. S'agreguen els resultats individuals dels càlculs en el resultat final

Una de les variants de Monte Carlo més usada a biologia és la basada en cadenes de Markov (MCMC) i més en concret les que fan servir l'algoritme de Metropolis-Hastings, que usen moviments aleatoris per explorar l'espai (*random walk*) de cerca per conèixer la densitat de l'espai real (Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, & E. Teller, 1953). El mètode va generant configuracions al atzar que són acceptades o no segons un criteri probabilístic basat en la diferencia d'energia entre la configuració de partida i d'arribada. El test garanteix que per un mostreig prou llarg la

densitat de estats obtinguda reprodueix l'esperable en una distribució de Boltzman per una temperatura donada.

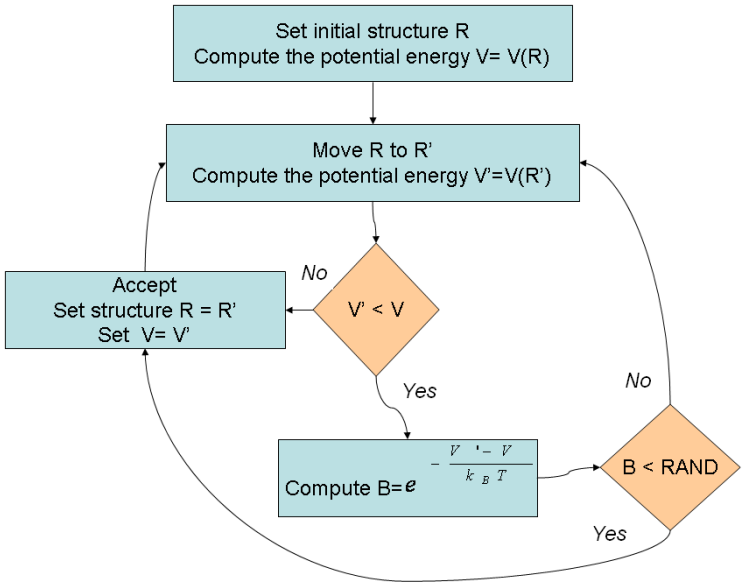

Figura 4.11 Esquema d'una simulació de MC Metropolis en sun sistema físic seguint la distribució de Boltzmann. RAND és un número aleatori generat en cada cicle.

Per sistemes simples, les modificacions en la estructura s'ajusten de manera que al voltant d'un 40%-50% de les conformacions són acceptades en el que s'ha vist que es millora la capacitat de mostreig. El canvi energètic es pot estimar a partir del càlcul de la energia potencial de la estructura inicial i la final. En una representació ideal de l'ADN a partir de 6 paràmetres helicoïdals (veure capítol estructura de l'ADN) la estimació de l'energia potencial d'una configuració és determinar a partir de la energia calculada usant les posicions en equilibri de cada paràmetre i les constants de força de cada propietat seguint un model harmònic (Pérez et al., 2007), veure equació 4.13

$$E = \sum_{steps} \Theta \cdot X_h \qquad (4.13)$$

A on la sumatòria corre per tots els passos considerats (steps), Θ es la matriu de rigidesa de mida 6x6 que representa les constants de força pures (diagonal) i

d'acoblament (fora de la diagonal) en l'espai helicoïdal (roll, twist, tilt, slide, shift & rise) y X_h es el quadrat de la diferencia entre el valor actual del pas (en l'espai helicoïdal) i el que te en l'equilibri. Com ja s'ha comentat en un altre capítol, la matriu de rigidesa surt a partir de la inversió de la matriu de covariança en l'espai helicoïdal obtinguda a partir de unes dinàmiques prèvies.

4.1.3.3 Recollida de dades: dinàmica essencial

Tant la simulació de MD com el MC complementen la informació experimental (estructures estàtiques d'ADN) amb informació sobre la funcionalitat de les biomolècules (flexibilitat). Però aquesta informació complementària, si no es processada, rarament aportarà resultats significatius. Existeixen diferents anàlisis que és poden fer a partir d'un mostreig d'equilibri de l'ADN (Noy, 2008; Manuel Rueda, 2005) obtingut per tècniques de MC, o més comunament MD:

- El RMSd o la desviació quadràtica mitjana respecte a una estructura referència (experimental, inicial, promig, ...). És una mesura àmpliament usada com a referent de estabilitat i convergència general d'una trajectòria.

- Informació estructural mitjana. En el cas dels àcids nucleics és possible calcular una sèrie de paràmetres estructurals promig com per exemple, els paràmetres helicoïdals. També és aplicable a les pautes de distància atòmiques per verificar o identificar interaccions de ponts d'hidrogen.

- Anàlisis de flexibilitat. La informació extreta de la MD o el MC es pot usar per calcular les propietats d'elasticitat dels paràmetres dels àcids nucleics assignats al mateix temps una mesura energètica als diferents graus de llibertat. Això pot aplicar-se als paràmetres helicoïdals usant una aproximació quasi-harmònica com la desenvolupada per Lankas (veure més adalt i capítol anterior).

- El potencial d'interacció molecular. És una mesura del perfil d'interacció d'una

molècula. En ell es calcula el potencial d'interacció clàssic entre un àtom de prova i la estructura promig de la simulació (J L Gelpí et al., 2001)

- Dinàmica essencial. La trajectòria d'una simulació pot resultar una mica caòtica, però molts cops oculta patrons en el tipus de moviment que segueix la molècula. Aquests patrons poden ser considerats els moviments essencials de la molècula.

L'últim tipus d'anàlisi és es basa en extreure els moviments importants d'una dinàmica no es trivial i requereix l'ús de potents eines estadístiques, destacant-ne l'anàlisi de components principals.

L'anàlisi de components principals (PCA, en anglès *principal component analysis*) és una tècnica estadística que permet eliminar la redundància de les dades i així reduir la dimensionalitat del sistema. Consisteix en una transformació lineal ortogonal de canvi de base. Per aquesta raó, es diagonalitza la matriu de covariàncies obtinguda en la MD o el MC buscant la direcció que expliqui més variabilitat i que d'aquesta forma contingui més informació (Noy, 2008; Manuel Rueda, 2005).

En la majoria de macromolècules s'ha vist que gran part de la variabilitat pot explicar-se amb relativament pocs graus de llibertat (pocs moviments essencials) que impliquen simultàniament a la majoria d'àtoms (Noy, 2008; Manuel Rueda, 2005). Els moviments essencials representen el sub-espai que conté la informació principal sobre la flexibilitat de les configuracions possibles dels àcids nucleics.

4.2 BASES DE DADES BIOINFORMÀTIQUES

La finalitat última de moltes de les eines bioinformàtiques exposades en aquest capítol és la d'aportar a la comunitat científica un coneixement en l'àmbit de la biologia. Perquè les eines en si siguin útils, és un política ben acceptada la de posar-les a disposició pública. Però de vegades més interessant que les eines desenvolupades són els resultats en si mateixos que han d'estar també accessibles a la comunitat en forma

d'una base de dades oberta i de fàcil de manipulació.

Les bases de dades han estat una de les claus de l'èxit dels recents avenços en la biologia molecular. Un exemple molt representatiu és la publicació del genoma humà (J C Venter et al., 2001) que posa de relleu les següents conclusions:

- Els resultats poden tenir més impacte que les eines bioinformàtiques usades.

- És essencial cuidar l'accés al contingut per a que aquest sigui usat amb èxit. La forma més simple (i menys eficient) d'accés és un repositori FTP de fitxers, una més avançada seria un servidor SQL i la més sofisticada (com en el cas del genoma humà) una eina especialment desenvolupada per explorar el seu contingut.

- Les dades són dinàmiques i evolucionen amb el temps pel que cal preveure mecanismes per manejar aquest element (p.ex. versions de la compilació).

- És difícil que una base de dades tingui èxit si no es capaç de combinar-se amb altres. En el cas del genoma humà, les dades de la seqüència van ser publicades de forma conjunta amb altres bases de dades com ara els gens, mutacions, etc.

Tot i que a vegades és difícil fer una classificació entre les bases de dades, ja que molts cops entre elles comparteixen informació, existeix encara una divisió entre llocs de contingut sobre ADN i llocs sobre contingut de proteïnes. Evidentment existeixen excepcions, llocs on la informació disponible difícilment podria ser classificada en alguna d'aquestes dues seccions. És el cas del Mowserv ("Mowserver") repositori de l'Institut Nacional de Bioinformàtica ("Spainsh National Institute of Bioinformatics") d'eines bioinformàtiques. L'esforç d'aquesta institució no és només posar en coneixement algoritmes bioinformàtics, sinó que també posa a disposició de l'usuari un entorn on les eines poden ser executades de forma individual o conjunta a través del protocol BioMoby ("BioMoby").

La quantitat de bases de dades del camp de la bioinformàtica disponibles avui en dia és enorme, i creix cada any. Revistes especialitzades dediquen espais especials a

recopilar aquelles que poden tenir més impacte en el camp. És el cas del suplement anual *Database Issue* de la revista *Nucelic Àcids Research*, on es poden trobar totes les bases de dades introduïdes en aquesta memòria. A continuació es descriuen les bases de dades que han intervingut més directament en el desenvolupament del treball presentat.

4.2.1 Bases de dades d'ADN

Les bases de dades d'ADN giren la majoria d'elles entorn de les seqüències genòmiques i principalment del genoma humà. Els treballs publicats amb el primer seqüenciament del genoma humà (J C Venter et al., 2001) van ser depositats a la vegada que un conjunt de bases de dades associat, en la seva majoria anotacions de seqüència. Tres de les primeres bases de dades introduïdes (ENSEMBL, NCBI, Genome Browser) donen una visió global, però complerta de genomes seqüenciats. DBTSS i CAGE TSS són dues col·leccions de TSS generades a partir de mètodes experimentals.

4.2.1.1 NCBI

El National Cententer for Biotechnology Information (NCBI) és una institució fundada fa dues dècades que a part de crear una de les bases biològica més importants a nivell mundial també s'encarrega de desenvolupar programari per analitzar el genoma, i té les seves pròpies línies de recerca en la biotecnologia. El seu objectiu és millorar la comprensió dels processos moleculars que afecten la salut humana. La seva base de dades està de fet composada per desenes de bases de dades inter-conectades, però independents la principal de les quals és el banc de genomes seqüenciats amb més de 25 organismes eucariotes superiors (incloent l'home), centenars de organismes microbials i milers de virus (D. L. Wheeler et al., 2008). A part del banc de genomes l'NCBI disposa de bases de dades sobre referències bibliogràfiques, de gens, de taxonomia d'organismes, homologia, patologies, SNPs o dominis de proteïnes entre moltes altres (D. L. Wheeler et al., 2008), i totes elles disponibles en http://www.ncbi.nlm.nih.gov/. Tres d'aquestes bases de dades, detallades a continuació han tingut rellevància en el desenvolupament de la següent memòria.

4.2.1.1.1 Refseq

La base de dades RefSeq (*Reference Sequence*) proveeix una col·lecció de referències a mRNAs, proteïnes i regions genòmiques curada manualment, i a més és complementa amb seqüències derivades computacionalment de nucleòtids i aminoàcids (Pruitt, Tatusova, & Maglott, 2005). La base de dades RefSeq està disponible a través d'un repositori de fitxers (FTP), però es pot accedir a través de la seva informació per la interfície web del NCBI (http://www.ncbi.nlm.nih.gov/RefSeq/) o usant la eines d'alineament de seqüència com BLAST (Jian Ye et al., 2006). El nombre de seqüències en RefSeq ha crescut un 33% en el darrer any i actualment està disponible la versió 22, amb 6.1 milions de seqüències de 4500 organismes diferents (Pruitt et al., 2005).

4.2.1.1.2 dbSNP

La bases de dades dbSNP és un repositori de mutacions d'un sol nucleòtids o de petites eliminacions/insercions (SNPs). Està disponible a través de FTP o de la web http://www.ncbi.nlm.nih.gov/projects/SNP/ i conté més de 12 milions de SNPs només pel genoma humà (S T Sherry et al., 2001). La bases de dades dbSNP no només dóna informació sobre la localització i les seqüències alternatives de cada mutació, també proporciona dades sobre quina és la freqüència de cada al·lel o quin és el estatus de la validació (validat experimentalment, localitzat computacionalment, ...). a part de descarregar-se la col·lecció sencera de SNPs en un genoma o gen, la bases de dades permet crear informes en XML sobre genotips.

4.2.1.1.3 Omim

La base de dades Omim (*Online Mendelian Inheritance in Man*) és de fet una versió electrònica del catàleg (amb el mateix nom) de gens i desordres genètics humans editat per la Universitat Hopkins (Hamosh, A. F. Scott, Amberger, Bocchini, & Victor A. McKusick, 2005). La base de dades conté informació sobre fenotips o gens associats a patologies, els polimorfismes, incloent una extensa descripció de la patologia, els noms i la localització dels gens, el patró d'herència i una bibliografia detallada. Actualment la base de dades compta amb aproximadament 18000 entrades que són públicament accessibles a través de la web http://www.ncbi.nlm.nih.gov/omim. Malauradament,

al ser aquest cas d'una col·lecció creada a partir d'un catàleg en suport físic, no és caracteritza per tenir una interfície massa eficient.

4.2.1.2 Ensembl

Ensembl (Ewan Birney et al., 2004) és un projecte comú entre el Laboratori de Biologia Molecular Europeu (EMBL) i de l'Institut Sanger per desenvolupar una plataforma o sistema informàtic que produeixi una anotació automàtica dels genomes eucariotes escollits. Ensembl és un projecte bioinformàtic per organitzar informació al voltant de genomes llargs (entre ells l'humà).

Ensembl és completament lliure per a totes les dades i programari. Aquest és accessible a través d'una web interactiva (http://www.ensembl.org/), a través d'un repositori de fitxers, pel protocol SQL o a partir d'un potent eina per minar de dades: BioMart.

4.2.1.2.1 Biomart

BioMart, originalment conegut com EnsMart (Kasprzyk et al., 2004), és una sistema de maneig de dades orientat a resoldre preguntes formulades per l'usuari, desenvolupat conjuntament per d'Institut Europeu de Bioinformàtica i el d'Institut de Recerca en Càncer d'Ontario. El sistema de Biomart pot ser usat per qualsevol tipus de dades i és especialment indicat per minar dades tal com fer cerques complexes d'informació. BioMart és presenta com a una eina amb interfície web (http://www.biomart.org/) o com un paquet instal·lable i configurable en el propi ordenador. L'aplicació no només dona accés a dades biològiques de Ensembl, també permet accedir a aplicacions de text i gràfiques i és programable a través de serveis-web i interfícies obertes en llenguatges de programació Perl o Java. El programa també facilita la transformació de dades en un format comprensible per BioMart. Seguint la filosofia de Ensembl, Biomart és codi obert (*open source*).

4.2.1.3 Genome Browser Database

El Genome Browser Database (GBD) de la Universitat de California Santa Cruz ofereix una col·lecció de anotacions genòmiques incloent conservació entre espècies,

gens, prediccions de gens, mRNAs, fragments de mRNAs alineats al genoma, expressió i regulació de gens, SNPs i compilacions de genomes, tot accessible a traves d'eines de visualització, comparació i anàlisis disponibles en http://genome.ucsc.edu. El GBD s'ha centrat sempre principalment en genomes de vertebrats, especialment l'home, pel qui disposa també la informació completa de ENCODE (T. E. P. Consortium, 2004).

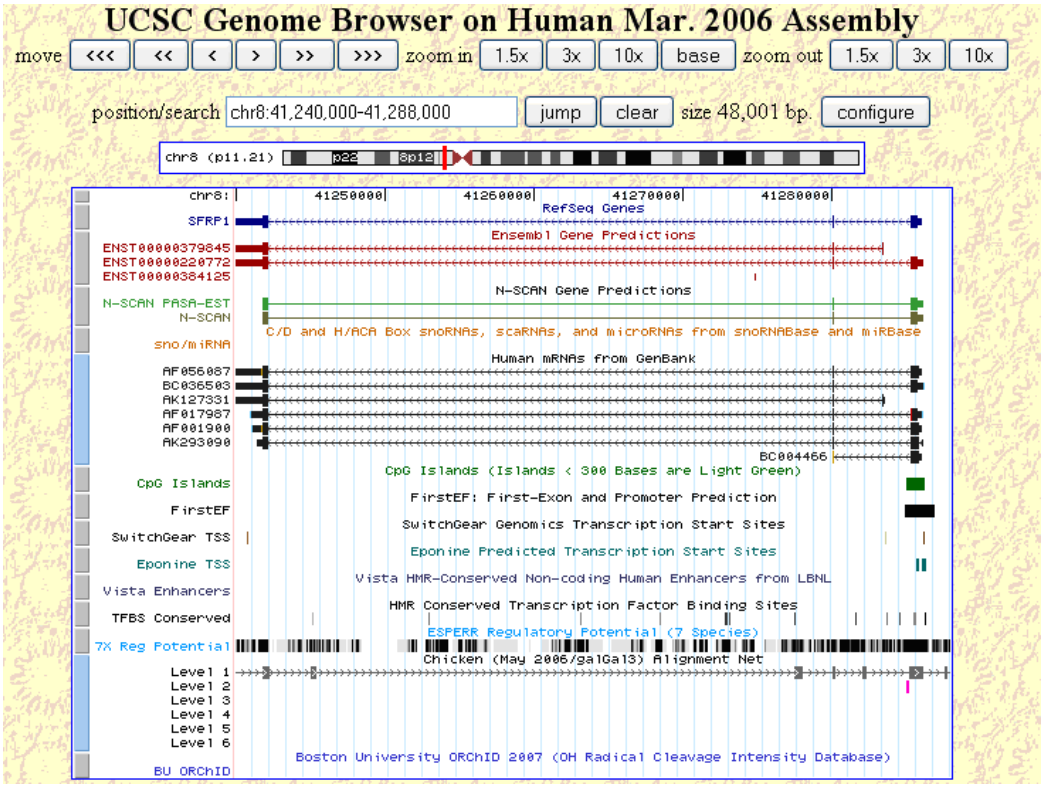

Figura 4.12 Vista del navegador del genoma disponible a la base de dades GBD

Les dades del GBD són accessibles a través d'un repositori de fitxers, del protocol SQL o per una interfície semblant a BioMart anomenada Table Browser (Karolchik et al., 2004). La eina més usada per explorar les anotacions genòmiques és el navegador del genoma (*Genome Browser*, veure figura 4.12) que permet fer zoom i desplaçar-se per sobre dels cromosomes, mostrant els resultats del treball global d'anotacions en la seqüència. Aquí la eina Blat (W. James Kent, 2002) juga també un paper fonamental, ja

que permet localitzar ràpidament seqüències en el genoma.

4.2.1.4 Dbtss

La Dbtss és una bases de dades de TSS anotats en els genomes de l'home i el ratolí (Wakaguri, Yamashita, Yutaka Suzuki, Sugano, & Nakai, 2008). La col·lecció de Dbtss és única, ja que els TSS han estat determinats experimentalment usant els seqüenciadors d'última generació. Les dades de Dbtss són accessibles a través de la web http://dbtss.hgc.jp/. A més, per proveir a l'usuari de vies per interpretar una font de dades tant massiva, el portal té disponibles dues eines analítiques: la primera per connectar informació d'expressió amb llocs d'unió a factors de transcripció i la segona un navegador per examinar la conservació entre especies de promotors i transcrits. La primera versió de Dbtss va sortir al 2002 i la última versió, la sisena que data del setembre del 2006, disposa de 19 milions de terminals 5' de seqüències (Wakaguri et al., 2008).

4.2.1.5 Cage TSS

La tècnica *cap-analysis gene expression* (CAGE) es basa en la preparació i seqüènciament dels 20 nucleòtids inicials de la regió 5' del ARNm. Aquetsa es una tècnica d'alt rendiment per l'anàlisi de la expressió gènica i la l'estudi de TSS que s'ha usat per predir de forma massiva inicis de transcripció en home i ratolí (Carninci et al., 2005, 2006). Aquestes dades es van posar a disposició pública en la web http://gerg01.gsc.riken.jp/cage/. Més que una base de dades aquesta web s'encarrega de emmagatzemar els resultats del treball Carninci et al., 2005. L'accés a la informació no és fàcil, però afortunadament es poden consultar les dades de la regió Encode a través del GBD.

4.2.2 Bases de dades de Proteïnes

Les bases de dades de proteïnes són més antigues i, per tant tenen generalment, més tradició que les d'ADN. Existeixen bases de dades orientades a acumular el màxim d'informació sobre les proteïnes conegudes; com per exemple, Uniprot o TrEBML i

altres centrades en aspectes concrets d'elles; com és el cas de PDB o GO que tenen informació sobre estructura i funció respectivament. Algunes bases son genèriques, com les anteriors i altres estan dedicades específicament a algunes famílies de proteïnes, com es el cas de TRANSFAC (factors de transcripció).

4.2.2.1 Uniprot

La *Universal Protein Resource* (UniProt) és un recurs per la comprensió de la seqüència i anotacions de proteïnes disponible des de http://www.uniprot.org/. És fruit de la col·laboració entre l'Institut Europeu de Bioinformàtica (EBI), l'Institut Suís de Bioinformàtica (SIB) i la *Protein Information Resource* (PIR). Fins fa uns anys el EBI i el SIB produïen les bases de dades Swissport i Trembl, mentre la PIR tenia la Pirpsd. Al 2002 els tres instituts van decidir coordinar els recursos i unir el seu coneixement per formar el consorci Uniprot. La força de la base i popularitat de la base de dades, però bé principalment de Swissprot.

4.2.2.1.1 Uniprot/Swissprot

Uniprot/Swissprot és la base de dades de seqüència de proteïnes més antigues (desde 1983) i popular. Es pot accedir de forma independent i alternativa al seu contingut a través de la web del SIB: http://www.expasy.ch/sprot/. Cada entrada fa referència a un proteïna i conté apart de la informació principal (la seqüència, referències bibliogràfiques, ...) anotacions (i enllaços a anotacions d'altres bases de dades) sobre:

- Funció de la proteïna
- Modificacions post-transcripcionals (fosforilització, acetilització, ...)
- Dominis de proteïnes i llocs d'interacció (unió d'ATP, ...)
- Estructura secundària de la proteïna (helix-alfa, fulles-beta,...)
- Homologia amb altres proteïnes

- Patologies associades amb deficiències de la proteïna (enllaç amb OMIM)

- Variacions amb la seqüència (SNPs)

- Informació sobre les isoformes, estroncament alternatiu (Boeckmann et al., 2005)

Uniprot/Swissprot es distingeix d'altres bases de dades de seqüència de proteïnes per tres grans criteris:

1. L'anotació de la seqüència de les proteïnes i del coneixement associat, es fa de manera manual i s'actualitza periòdicament per cada proteïna individualment. L'origen de les dades sol venir de publicacions científiques, revisions regulars de famílies o grups de proteïnes o directament de opinions d'experts en el camp.

2. Un dels problemes principals de les bases de dades és la redundància. Aquest origen està en les incompatibilitats entre versions de les bases de dades o per l'origen distribuït de la informació (dos grups independents poden estar enviant una informació no idèntica, però altament redundant). Uniprot/Swisprot fa un esforç per acumular productes d'un mateix gen en una sola entrada.

3. Moltes de les anotacions sobre les proteïnes de la bases de dades són de fet accessos a altres bases de dades. De forma inversa Uniprot/Swisprot està interconectada en altres bases de dades; com per exemple, el NCBI, Ensemble, PDB (veure en seccions següents), etc. Actualment Uniprot/Swisprot esta inter-relacionada amb 60 bases de dades.

4.2.2.1.2 Uniprot/Trembl

Uniprot/Trembl és la versió no-curada de Uniprot/Swissprot. Conté la traducció de totes les seqüències codificants presents en Ensembl i les proteïnes de UniProtKB/Swiss-Prot. La base de dades està enriquida amb anotacions i classificacions automàtiques. Trembl és molt més recent que Swissprtot i va néixer com una alternativa més àgil a aquesta darrera. Donat que apart de les entrades de

Swissprot, integra anotacions bioinformàtiques el volum de dades de Uniprot/Trembl supera de llarg Uniprot/Swissprot (veure figura 4.13), però les dades i anotacions són en general d'una qualitat inferior.

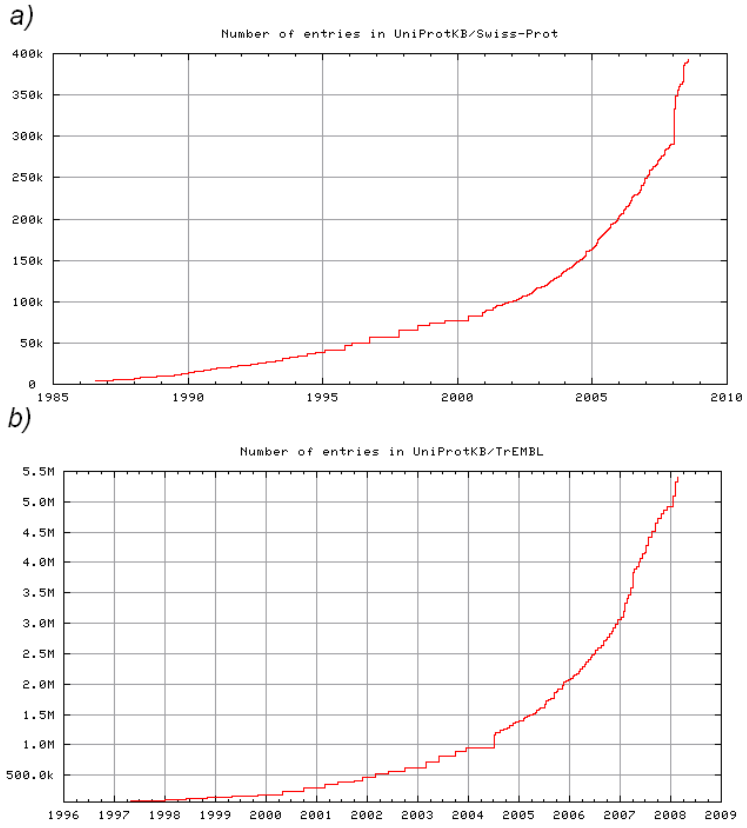

Figura 4.13 Creixement de les entrades de (a) Uniprot/Swissprot i (b) Uniprot/Trembl

4.2.2.2 PDB

El *Protein Data Bank* (PDB) no és estrictament una base de dades, sinó un banc de dades fundat el 1971 on els usuaris deixen informació sota la seva responsabilitat. PDB és un repositori d'estructures en tres dimensions d'estructures de macromolècules biològiques, determinades experimentalment (Dutta et al., 2008). Tot i que, la gran majoria de estructures correspon a proteïnes (senceres, parcials o complexes), també s'hi poden trobar àcids nucleics. Les dades en l'arxiu de PDB és lliure a través de la

web http://www.rcsb.org/pdb. PDB ha aconseguit ja arribar a les 50,000 estructures, doblant la mida des del 2004, fruit fonamentalment dels projectes massius de genòmica estructural.

4.2.2.3 GO

La *Gene Ontology* (GO) és un projecte (http://www.geneontology.org/) per proveir a la comunitat d'anotacions de gens d'un conjunt de vocabulari controlat i estructurat sobre els productes del gens (normalment proteïnes), que actualment compta amb 1300 termes. Les ontologies (veure figura 4.14) s'extreuen i es refinen de diferents àrees de la biologia imposant en un refinament posterior la seva estructura (G. O. Consortium, 2004; The Gene Ontology Consortium, 2008). Existeixen 3 tipus principals d'ontologies:

- Procés biològic: ontologies que fan referència a qualsevol procés específicament pertinent al funcionament de les unitats de vida com cèl·lules, teixits, òrgans i organismes. Un procés és una col·lecció d'esdeveniments molecular amb un principi i final definit.

- Funció molecular: fa referència a activitats (com per exemple, de catàlisi o d'unió) que descriuen accions del producte del gen a nivell molecular. Un producte de gen met exhibir una o més funcions moleculars.

- Component cel·lular: la part de la cèl·lula o l'entorn extracel·lular on el producte del gen es localitza. El producte d'un gen es pot localitzar en una o més parts de la cèl·lula.

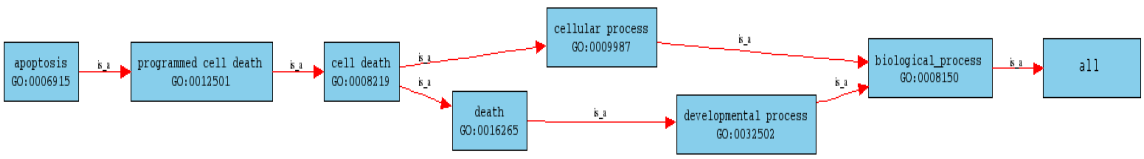

Figura 4.14 Ontologies sobre productes de gens en la bases de dades GO. La ontologia

"apoptosis" està dins el grup de processos biològics.

La base de dades proveeix no només de les ontologies, sinó també de les seves associacions amb les proteïnes. En aquest cas és útil conèixer a través de la base de dades la evidència d'aquesta associació, que pot ser:

- Evidencia experimental: per experiment, interacció física, interacció genètica, patró d'expressió entre d'altres tipus.

- Evidència computacional; com per exemple, de proteïna a proteïna per similitud estructural, ortologia de seqüència, context genòmic o per anàlisis computacionals.

- Evidencia basada: en afirmació d'autor, que pot ser reproduïble o no reproduïble.

- Evidencia curada.

- Altres evidencies; com per exemple, l'anotació per inferència electrònica.

Gene Ontoly disposa també del navegador Amigo http://amigo.geneontology.org), un eina-web per la cerca de termes GO i anotacions que ha estat orientada a simplificar i millorar la interfície amb l'usuari. Els usuaris avançats també tenen la opció d'accedir a la informació a través del protocol SQL.

4.2.2.4 Transfac

La base de dades Transfac és un producte comercial de la empresa Biobase (http://www.gene-regulation.com/pub/databases.html). Es tracta d'un paquet de coneixement sobre llocs d'unió de factors de transcripció. Alguns tipus d'ús i accés a la base de dades requereix el pagament d'una llicència (tant per entitats comercials com per acadèmia).

La unió dels TF al llocs d'ADN d'aquesta base de dades han estat provats experimentalment demostrant, també la capacitat de regulació sobre els gens. La gran compilació de llocs d'unió disponible permet una derivació de qualitat de PWM de TFBS.