UNIVERSITAT DE BARCELONA

Nous desenvolupaments, aplicacions bioanalítiques i validació de mètodes de resolució multivariant

Joaquim Jaumot Soler

Tesi Doctoral
20 de juny de 2006

Capítol 2

Química dels àcids nucleics

Durant segles, els éssers humans han observat el procés de l'herència sense entendre gaire bé com es transmetien els trets físics de pares a fills. Moltes cultures han fet servir aquestes observacions per millorar les seves condicions econòmiques, en camps com la ramaderia o l'agricultura [1]. La investigació de l'herència, que actualment s'anomena Genètica, no va començar fins al segle XIX. A principis del segle XX, els científics van començar a admetre de forma generalitzada que els trets físics s'hereten de forma discreta (que posteriorment es van anomenar gens) i que els cromosomes de l'interior del nucli de la cèl·lula són els dipòsits de la informació genètica. Finalment, es va elucidar la composició química dels cromosomes i es va identificar l'àcid desoxiribonucleic (ADN) com a portador de la informació genètica. Actualment, al conjunt complet de la informació genètica d'un organisme codificat en la seqüència de nucleòtids del seu ADN se l'anomena genoma [2].

La Biologia Molecular és la ciència que estudia l'estructura i la funció dels gens. El descobriment de l'estructura de l'ADN com un dúplex helicoïdal de polímers de nucleòtids per James Watson i Francis Crick el 1953 [3] ha permès als científics reexaminar la majoria dels fenòmens biològics. Així, en les últimes cinc dècades s'ha formulat una visió general de l'herència biològica i de la transferència d'informació genètica que ha permès establir l'anomenat "Dogma central de la Biologia Molecular" [4, 5], que descriu el flux d'informació genètica des de l'ADN fins a les proteïnes passant per l'ARN (Figura 2.1.):

1. La informació codificada en l'ADN consisteix en una seqüència específica de bases nitrogenades. El primer pas de la transmissió de la informació genètica consisteix en duplicar la cadena d'ADN original. Aquest procés s'anomena replicació, ja que s'obté una còpia o replicat de la informació genètica original.

2. El mecanisme de descodificació i utilització de la informació genètica pel govern dels processos cel·lulars comença amb la síntesi d'un altre tipus d'àcid nucleic, l'àcid ribonucleic (ARN). La síntesi de l'ARN es porta a terme mitjançant l'aparellament complementari de les bases de ribonucleòtids amb les bases d'una molècula de ADN en un procés també conegut com a transcripció.

3. A partir de l'ARN es sintetitzen les proteïnes i els enzims en un procés conegut com traducció.

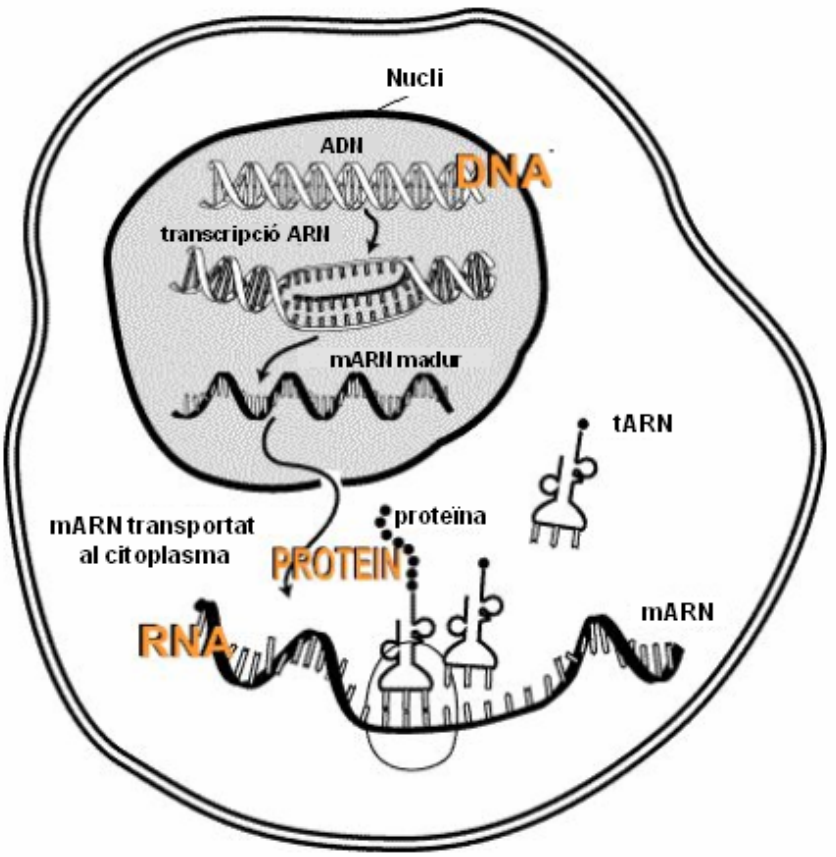


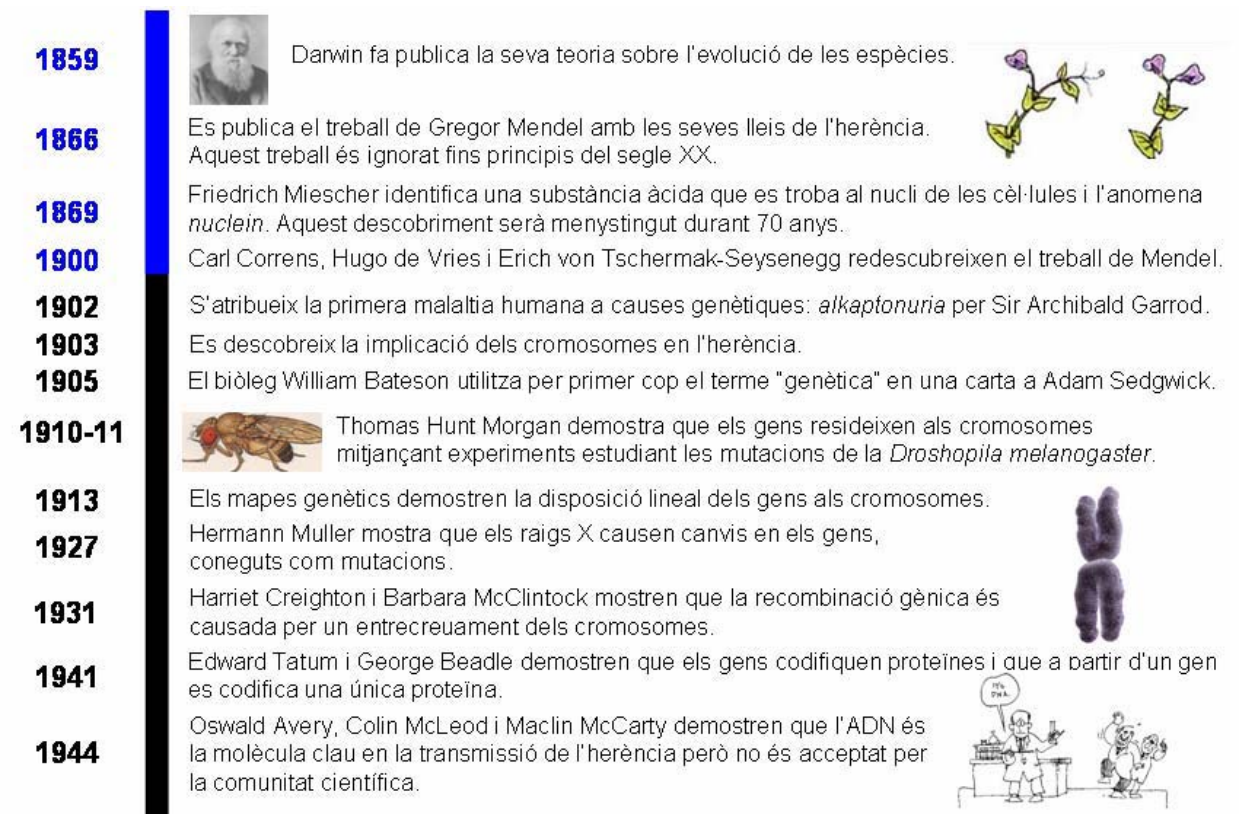
Figura 2.1. Representació gràfica del "Dogma central de la Biologia Molecular".

L'elucidació i l'anàlisi dels genomes complets de desenes d'espècies, com és el cas del genoma humà en el marc del projecte *Human Genome Project* (HGP)[6-8], han fet aparèixer una nova branca de la Biologia coneguda com Genòmica [9]. Aquesta nova branca de la Ciència intenta no únicament trobar l'estructura i la identitat dels gens en el éssers vius, sinó també entendre com funcionen de forma global les biomolècules a l'interior d'un organisme. A més, han aparegut altres branques afins com pot ser, per exemple, la Proteòmica [9]. La Proteòmica es pot definir com la Genòmica funcional a nivell de les proteïnes, és a dir, intentar correlacionar les proteïnes que es poden obtenir a partir d'un genoma.

El coneixement de l'estructura i les funcions dels gens ha permès l'aparició de tècniques que intenten aprofitar aquests nous coneixements per tal de poder satisfer

les necessitats humanes en camps tan variats com l'agrobiotecnologia, el medi ambient o la biomedicina. D'aquesta forma s'han desenvolupat experiments que permeten l'anàlisi simultani de milers de gens per tal de descobrir les seves funcions [10] o la utilització d'oligonucleòtids com a agents terapèutics [11]. Per exemple, dues teràpies que han proporcionat bons resultats són la teràpia antigènica i la teràpia antisentit. La primera es basa en el disseny d'oligonucleòtids capaços d'inhibir una part del procés de transcripció mitjançant la formació d'hèlices triples [12]. Això comporta el disseny d'oligonucleòtids amb una seqüència de bases adient ja que s'haurà d'enllaçar amb un fragment determinat de la cadena d'ADN. La teràpia antisentit es basa en el disseny d'oligonucleòtids capaços d'enllaçar-se a l'ARN missatger abans que es produeixi la traducció [13, 14].

L'estudi dels àcids nucleics com a molècula clau per a la transmissió de la informació genètica va començar a mitjans del segle XIX. A la Figura 2.2. es mostren algunes dates clau en la història de l'estudi dels àcids nucleics.



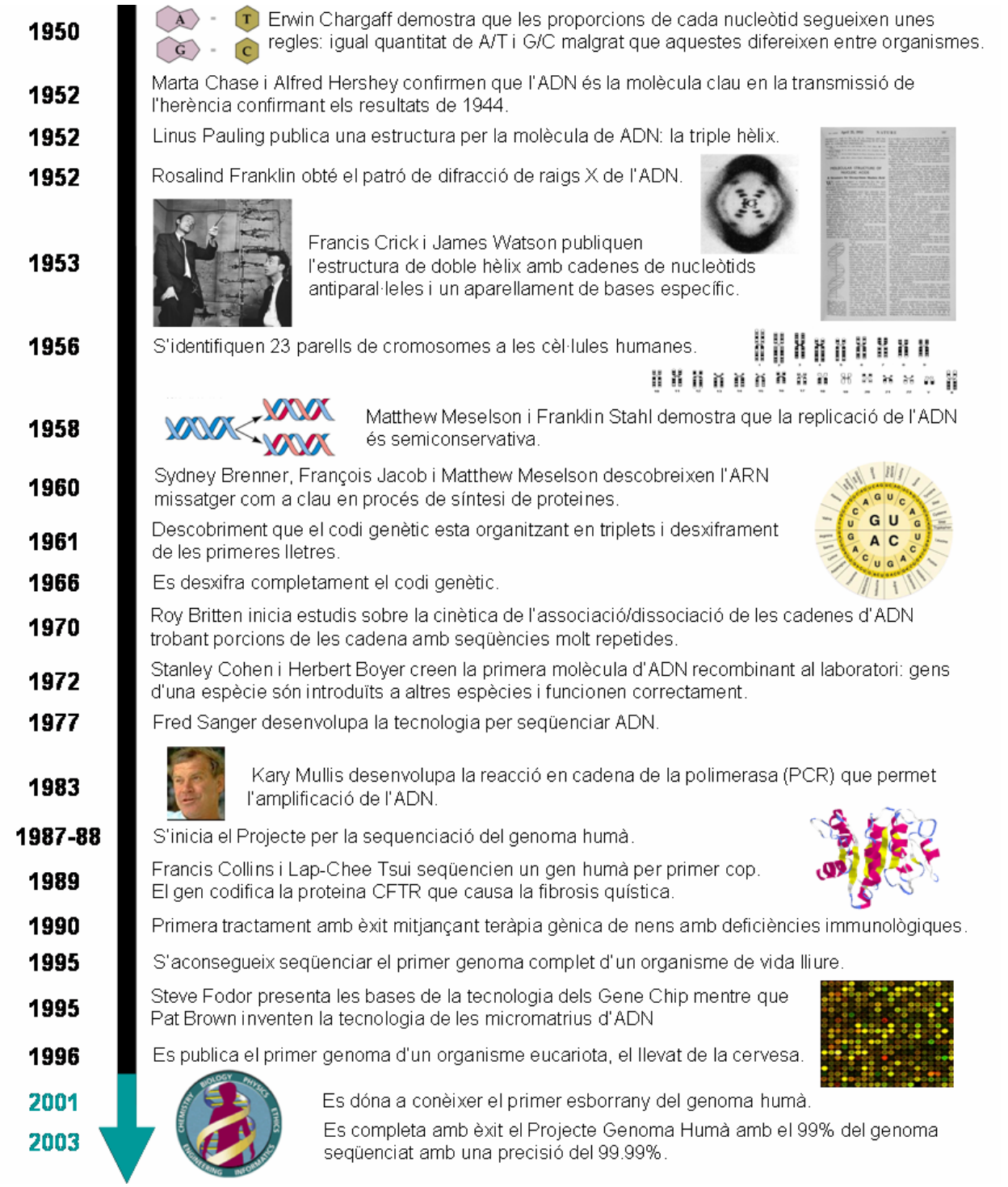


Figura 2.2. Cronologia del descobriments relacionats amb els àcids nucleics adaptat de [15].

2.1. Estructura dels àcids nucleics

2.1.1. Conceptes bàsics i estructures habituals

Existeixen dos tipus d'àcids nucleics, l'àcid desoxiribonucleic (ADN) i l'àcid ribonucleic (ARN). L'ADN i l'ARN poden considerar-se com polímers formats per quatre tipus de monòmers que reben el nom de nucleòtids [5]. Cada un d'aquests nucleòtids es pot considerar com el derivat 5'-monofosforilat d'un adducte sucre-base nitrogenada anomenat nucleòsid (Figura 2.3.) ja que són molècules de ribosa o desoxiribosa fosforilades amb bases puríniques o pirimidíniques unides als seus carbonis 1'. A les purines, aquest enllaç es produeix mitjançant el nitrogen 9, mentre que en les pirimidines l'enllaç es produeix mitjançant el nitrogen 1. L'enllaç entre el carboni 1' del sucre i el nitrogen de les bases s'anomena enllaç glicosídic. Donat que tots els àcids nucleics es poden considerar com polímers de nucleòtids se'ls pot assignar la denominació genèrica de polinucleòtids. Els polímers petits amb un nombre petit de bases reben el nom d'oligonucleòtids.

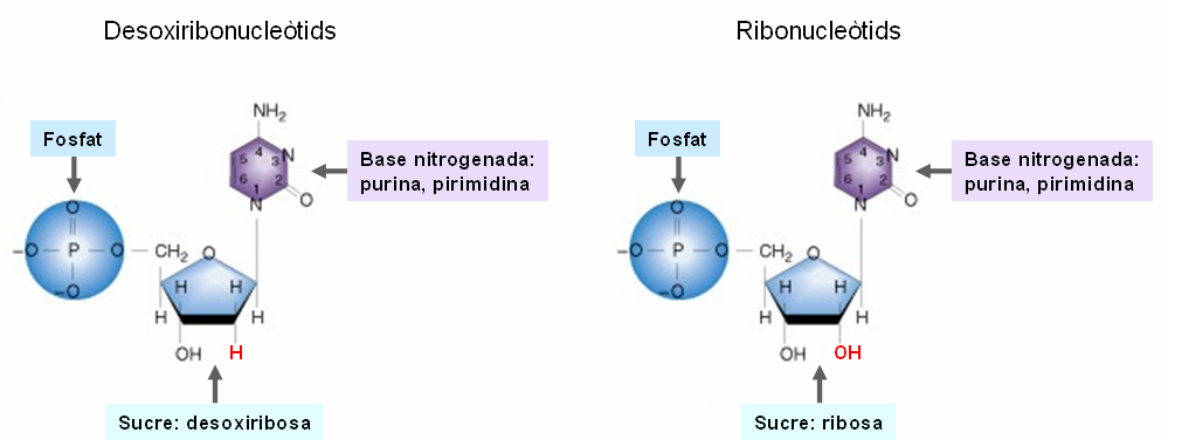


Figura 2.3. Estructura general de les unitats monomèriques dels desoxiribonucleòtids i dels ribonucleòtids.

Tant en el cas de l'ADN com de l'ARN, el nucleòtid conté un hidrat de carboni amb cinc àtoms de carboni, la ribosa pel ARN i la 2'-desoxirribosa pel DNA. La diferència entre els dos hidrats de carboni radica únicament el grup hidroxil 2' de la ribosa en l'ARN, que es troba substituït per un hidrogen a l'ADN. La connexió entre les successives unitats monomèriques als àcids nucleics es realitza mitjançant un grup fosfat unit a l'hidroxil del carboni 5' d'una unitat i a l'hidroxil 3' de la següent. Això produeix un

enllaç fosfodièster entre dos hidrats de carboni consecutius. D'aquesta forma es constitueixen cadenes llargues d'àcids nucleics que poden arribar a contenir milions d'unitats monomèriques. Per una altra banda, el grup fosfat és un àcid fort que va ser la causa per la qual, històricament, aquestes molècules es van anomenar àcids nucleics [9].

Els sucres units mitjançant l'enllaç fosfodièster constitueixen l'esquelet de la molècula d'àcid nucleic. Aquesta estructura repetitiva és incapaç de dur a terme la codificació de la informació genètica. La importància dels àcids nucleics en l'emmagatzematge i la transmissió de la informació radica en el fet que són heteropolímers, és a dir, en el fet que cada monòmer de la cadena conté una base heterocíclica que sempre va unida al carboni 1' de l'hidrat de carboni. Existeixen dos tipus de bases que es denominen purines i pirimidines. L'ADN té dues purines, adenina (A) i guanina (G), i dues pirimidines, citosina (C) i timina (T). L'ARN té les mateixes bases excepte que la timina es troba substituïda per l'uracil (U) (Figura 2.4.). Tanmateix existeix també una petita proporció de bases amb modificacions com ara la hipoxantina o la 5-metilcitosina [16].

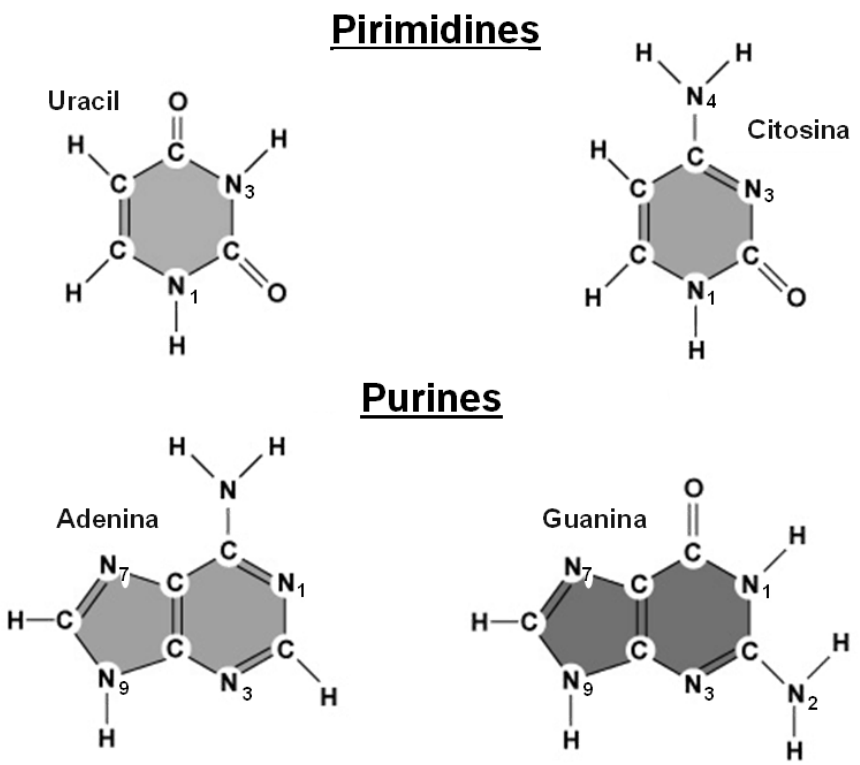


Figura 2.4. Estructura de les bases nitrogenades. A la meitat superior es troben les bases pirimidíniques i a la meitat inferior les bases puríniques constituents de l'ADN.

Quan es parla de l'estructura dels àcids nucleics s'ha de tenir en compte que la configuració d'un àcid nucleic es refereix als enllaços covalents presents a la molècula, mentre que la conformació es refereix a l'estructura tridimensional de la molècula [17]. Així, la configuració d'una molècula és constant mentre que la conformació depèn del medi en el qual es troba (temperatura, pH, concentració de sals, …). Això implica que mentre la conformació és dinàmica ja que una molècula pot adoptar diferents estructures a l'equilibri, la configuració és estàtica. D'aquesta forma, la configuració serà l'estructura primària de la cadena d'àcids nucleics mentre que la conformació serà l'estructura secundària o terciària. A continuació es mostra una descripció de cadascuna d'aquestes estructures.

Estructura primària

Les cadenes de polinucleòtids presenten entre d'altres dues característiques importants. En primer lloc, una cadena de polinucleòtids té un sentit o direccionalitat. L'enllaç fosfodièster entre les unitats monomèriques es produeix entre els carbonis 3' d'una unitat i el 5' de la següent. Així, els dos extrems d'una cadena són diferenciables ja que l'extrem 5' presenta un grup fosfat sense reaccionar i l'extrem 3' presenta un grup hidroxil també sense reaccionar. D'altra banda, una cadena de polinucleòtids té una individualitat proporcionada per la seqüència de les bases. Aquesta seqüència és el que es coneix com estructura primària o configuració. La importància de l'estructura primària és que la informació genètica s'emmagatzema en aquesta [1]. Un gen no és més que una seqüència concreta de bases d'ADN que codifica la informació mitjançant un llenguatge de quatre lletres, en el qual cada "lletra" és una de les bases nitrogenades.

Estructura secundària

Al plegar-se la cadena de nucleòtids es formen determinades disposicions localitzades dels nucleòtids adjacents que constitueixen el que s'anomena com estructura secundària.

L'elucidació de l'estructura secundària de l'ADN la van dur a terme James Watson i Francis Crick el 1953 [3]. Aquests van aprofitar una sèrie de descobriments duts a terme a l'inici de la dècada dels cinquanta com ara les lleis de Chargaff [18], les quals

postulen que el nombre de bases d'adenina és igual al nombre de bases de timina i que el nombre de bases de guanina és igual al nombre de bases de citosina, o la determinació de les formes tautomèriques cetòniques de les bases nitrogenades [19]. També va ser clau el desenvolupament de la tècnica de difracció de raigs X i les fotografies dels patrons de raigs X d'unes fibres d'ADN que Rosalind Franklin va proporcionar a Watson i Crick. A partir d'aquestes fotografies (Figura 2.5.) es va determinar que l'ADN presentava un patró creuat característic d'una estructura secundària helicoïdal [3].

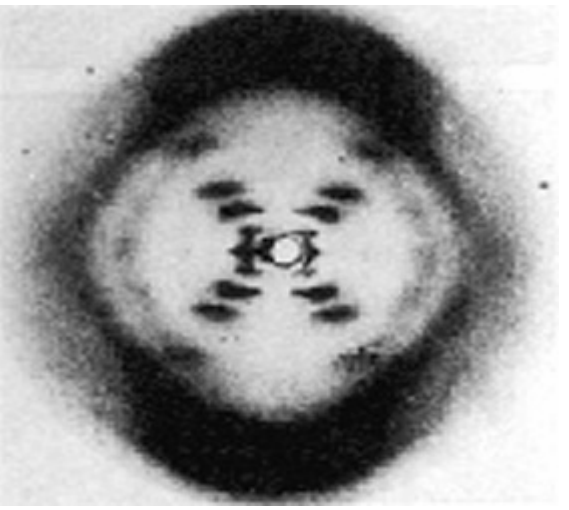


Figura 2.5. Patró de difracció de raigs X de la molècula d'ADN obtinguda per Franklin.

Les dades relatives a la densitat de la fibra van suggerir que hi havia d'haver dues cadenes d'ADN a cada molècula helicoïdal. El fet determinant per obtenir l'estructura correcta va ser la intuïció de Watson i Crick que una hèlix de doble cadena es podia estabilitzar mitjançant ponts d'hidrogen entre les bases de cadenes oposades si les bases s'aparellaven d'una manera concreta: els parells A·T i G·C que permeten l'aparició d'enllaços d'hidrogen forts entre les bases. La geometria d'aquests parells de bases permet que qualsevol seqüència de bases s'ajusti a la doble hèlix sense distorsionar-la i presenta dos ponts d'hidrogen en el cas del parell de bases A·T i tres ponts d'hidrogen en el cas del parell de bases G·C (Figura 2.6.). Aquest major nombre de ponts d'hidrogen proporciona als parells de bases G·C una lleugera major estabilitat [19].

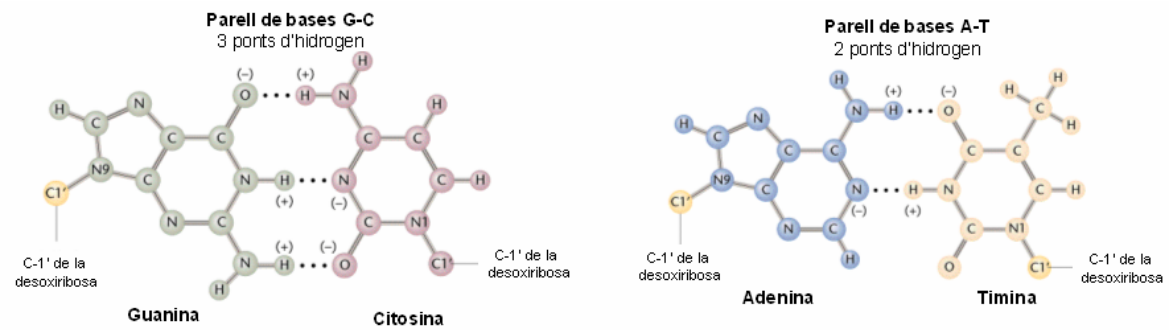


Figura 2.6. Aparellament tipus Watson i Crick de les bases G·C i A·T adaptat de [17, 20].

En el model proposat per Watson i Crick (Figura 2.7.), l'esquelet hidròfil de fosfat/desoxiribosa es situa a l'exterior, en contacte amb el medi aquós, mentre que els parells de bases s'apilen interiorment els uns sobre els altres, amb els seus plans quasi perpendiculars a l'eix de l'hèlix [3, 21]. Aquest apilament de les bases permet que es produeixin interaccions de van der Waals molt fortes entre elles, estabilitzant el conjunt de l'estructura. El model indica també que, encara que les bases es trobin a l'interior, s'hi pot accedir a elles a través de dos solcs espirals profunds, anomenats solc major (*major groove*) i solc menor (*minor groove*). El solc major proporciona un accés directe a les bases des de fora de l'hèlix mentre que el menor es troba davant de l'esquelet format pels sucres. Aquest fet implica que el model molecular de l'estructura d'ADN de cadena doble ha de tenir les dues cadenes d'ADN en sentits oposats. Donat l'aparellament de les bases A·T i G·C es pot dir que les cadenes de l'hèlix doble són complementàries i que, per tant, permeten l'autoreplicació, és a dir, que a partir de cadascuna de les cadenes de la doble hèlix es pot obtenir dues molècules d'ADN idèntiques a l'original [22, 23].

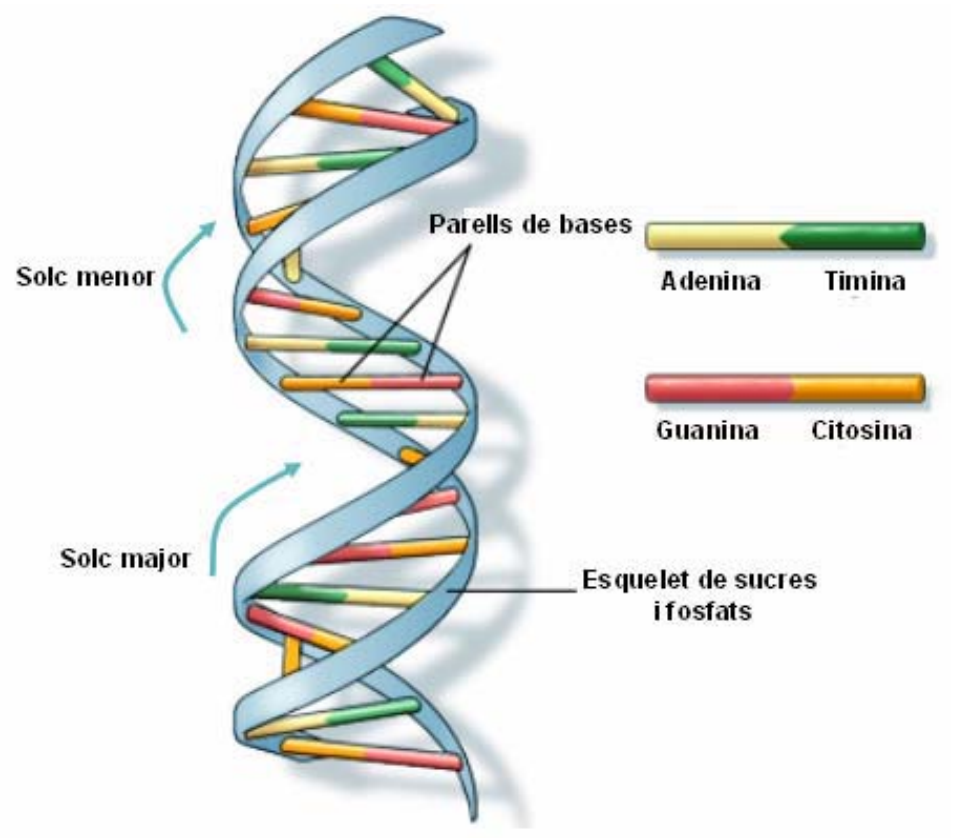


Figura 2.7. Model de la doble hèlix d'ADN proposat per Watson i Crick adaptat de [5].

En el moment en el que Watson i Crick van proposar el seu model, ja s'havien obtingut dos patrons de difracció de raigs X força diferents el que indicava que la molècula existeix en més d'una forma [24, 25]. La forma B, que s'observa a les fibres d'ADN preparades en condicions d'humitat elevada, i la denominada forma A, que s'observa en condicions d'humitat més baixa. Malgrat que l'hèlix B és la forma d'ADN que es troba a les cèl·lules, l'hèlix A també té una importància biològica ja que les molècules d'ARN de doble cadena i els híbrids ARN-ADN formen sempre l'estructura A [16].

Estructura terciària

L'estructura terciària dels àcids nucleics apareix quan existeix un plegament d'ordre superior dels elements d'una estructura secundària regular. En molts casos, les molècules d'ADN que es troben a les cèl·lules són circulars, és a dir, que no tenen extrems 5' o 3' lliures. Els cercles poden ser petits o grans i poden estar formats per una sola cadena o per dues cadenes en forma de doble hèlix a la seva forma B. Moltes d'aquestes molècules circulars es troben superenrotllades, presentant altres torsions addicionals al propi eix de l'hèlix [5].

2.1.2. Estructures secundàries del ADN poc habituals

Com s'ha esmentat a l'apartat anterior, les dues cadenes de la doble hèlix es mantenen unides degut a les interaccions per pont d'hidrogen que es produeixen entre els parells de bases nitrogenades.

Altres aparellaments de bases són també possibles malgrat que l'aparellament tipus Watson-Crick sigui el dominant en els àcids nucleics naturals. Aquests són els de Hoogsteen, el de *wobble* i les seves variants invertides: Watson-Crick invertit, Hoogsteen invertit i *wobble* invertit (Figura 2.8.) [17].

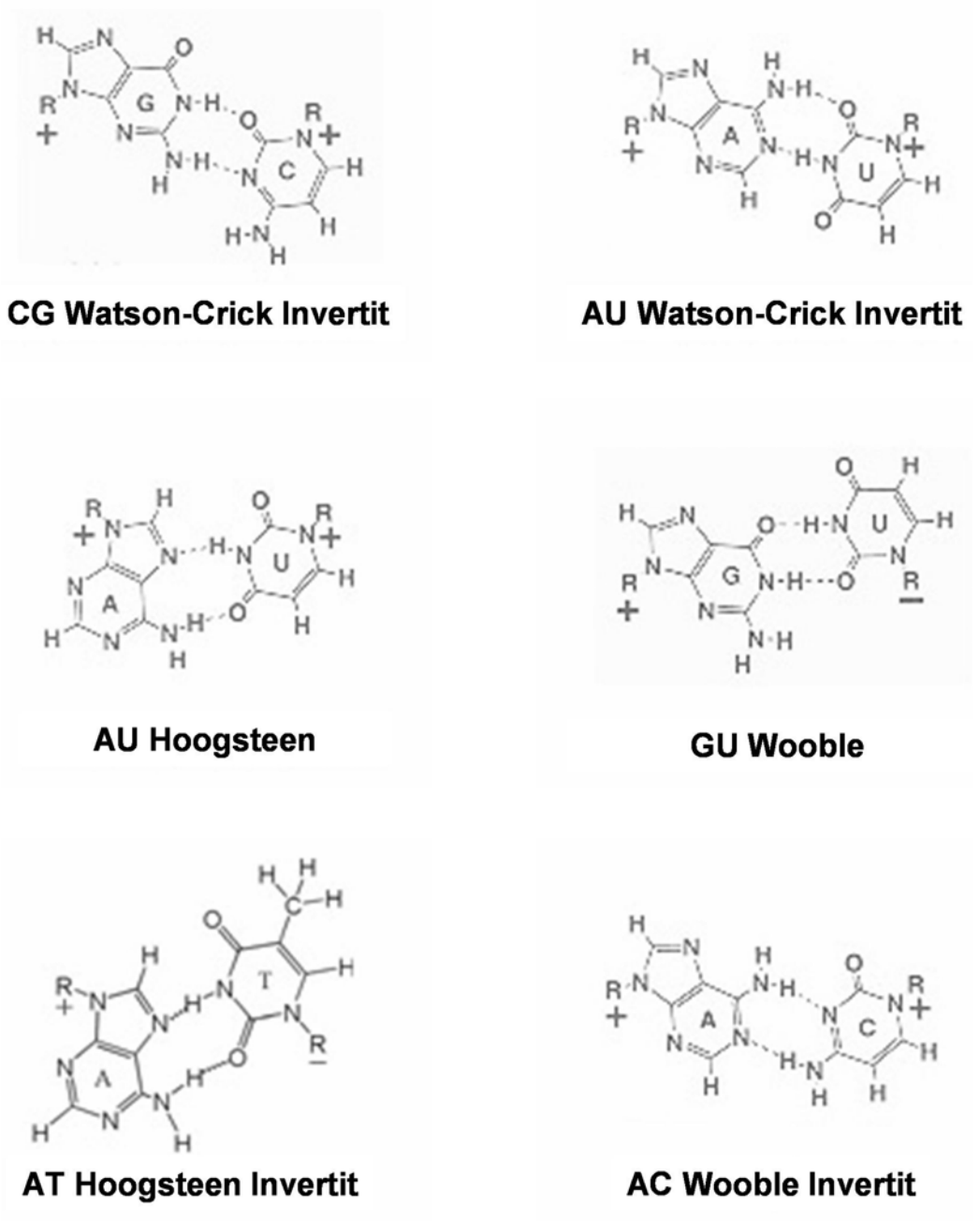


Figura 2.8. Altres aparellaments de bases: WC invertit, Hoogsteen, Hoogsteen invertit, *wobble* i *wobble* invertit adaptat de [17].

A la Figura 2.9. es mostren combinacions coplanars on poden intervenir més de dues bases. Els aparellaments de tres bases permetran la formació d'estructures triples. Això succeeix quan la cadena doble s'uneix a una tercera cadena mitjançant un aparellament tipus Hoogsteen.

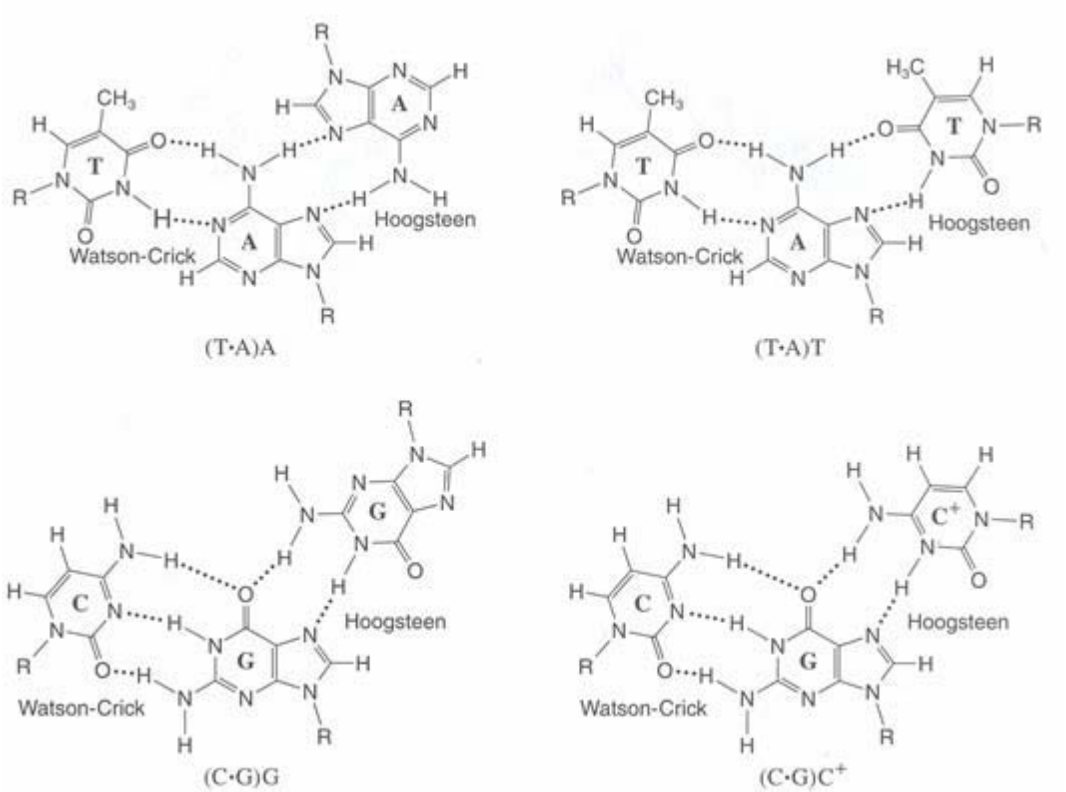


Figura 2.9. Aparellaments de tríades de bases (T·A)*A, (T·A)*T, (C·G)*G i (C·G)*C^+ adaptat de [17].

Finalment, s'han observat l'aparellament de quatre bases en el mateix pla. El que presenta més rellevància biològica és el que es coneix com *G-quadruplex* (Figura 2.10.) on es troben quatre guanines estabilitzades per ponts d'hidrogen i que està present als telòmers (extrems físics dels cromosomes eucariotes).

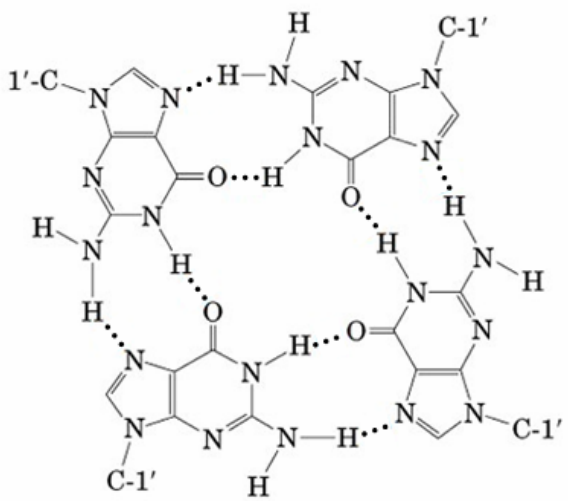


Figura 2.10. Aparellaments de 4 bases formant un *G-quadruplex* adaptat de [17].

Aquesta gran varietat de possibles aparellaments de bases provoca que els àcids nucleics (tant els oligonucleòtids com els polinucleòtids) puguin presentar múltiples motius estructurals [20].

- Regions d'una única cadena

Les molècules de nucleòtids d'una única cadena poden adoptar diverses estructures que depenen de la seva seqüència i de les condicions de la solució. A temperatura elevada o presència de substàncies desnaturalitzants, la majoria es trobaran en forma desestructurada o de *random coil* [5]. Aquesta estructura es caracteritza per la flexibilitat i la llibertat de rotació al voltant dels enllaços de l'esquelet la qual cosa permet canvis continus. En canvi, en condicions properes a les fisiològiques les interaccions d'apilament (*base stacking*) tendiran a formar regions d'hèlix d'una cadena de bases apilades.

- Estructures dobles

A més de les formes de B- i A- ADN descrites anteriorment s'ha descobert altres variacions de l'estructura de l'ADN depenent de les condicions del medi com poden ser el C-, el D- i el T-ADN. Però s'ha de destacar una altra estructura que presenta dues cadenes i forma helicoïdal però en la qual el gir es produeix cap a l'esquerra que es va anomenar Z-ADN i va ser descoberta l'any 1979 [16].

- Forquetes (*hairpins*)

Les estructures en forma de forqueta apareixen quan la cadena de l'àcid nucleic gira sobre si mateixa per tal de formar parells de bases, de forma que poden quedar regions de bases nitrogenades sense aparellar. Aquestes estructures són molt habituals en les cadenes d'ARN *in vivo* (ARN de transferència i ARN ribosòmic) i en les cadenes d'ADN apareixen alternades amb l'estructura de doble hèlix [16]. A la Figura 2.11. es representa en format tridimensional una estructura tipus forqueta.

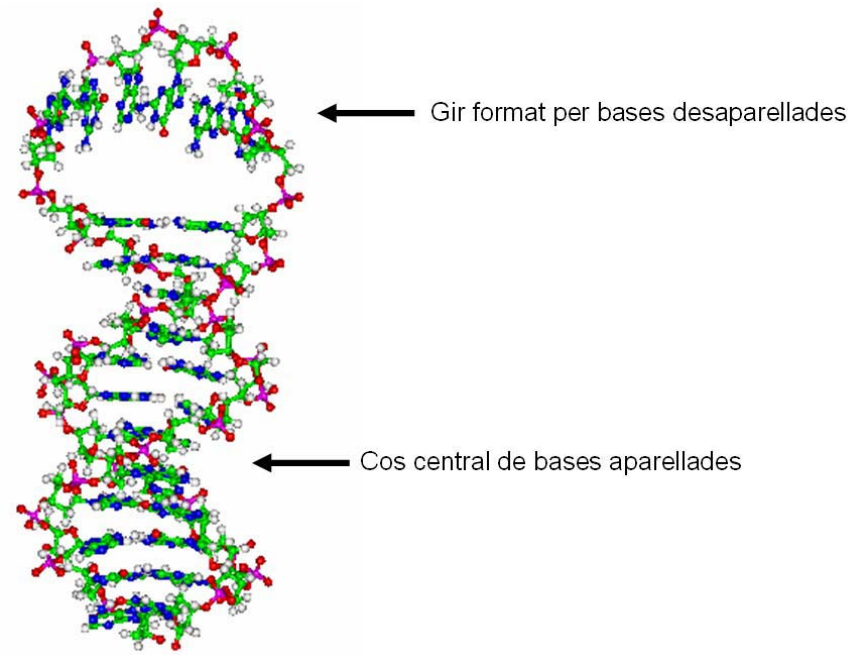


Figura 2.11. Estructura tipus forqueta amb un cos central amb parells de bases i un gir amb bases desaparellades adaptat de [26] .

Els oligonucleòtids amb estructura *dumbbell* consisteixen en una forqueta cíclica [27].

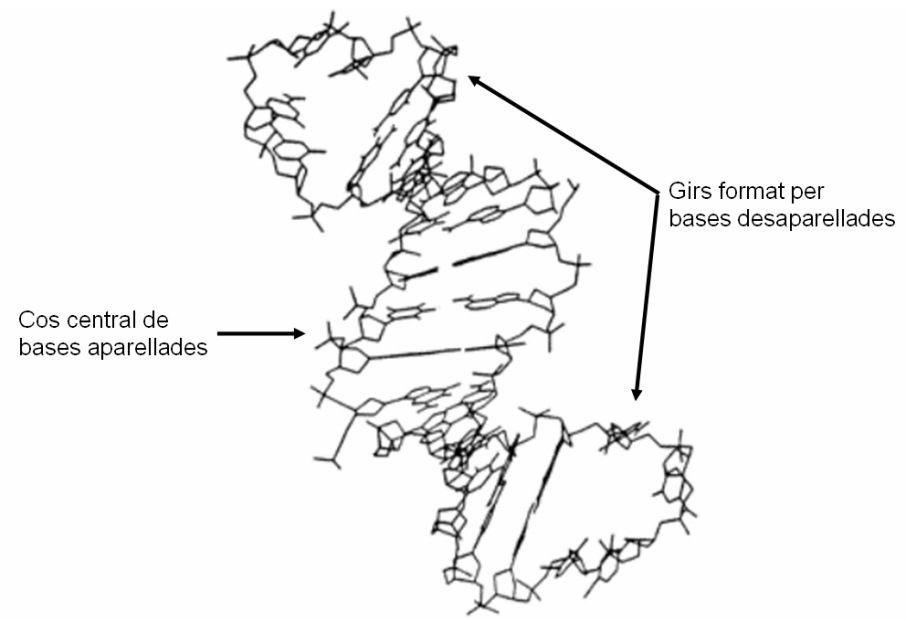


Figura 2.12. Oligonucleòtid cíclic amb una estructura tipus dumbbell [28].

A la Figura 2.12. es representa l'estructura tridimensional d'un oligonucleòtid en la conformació *dumbbell*. S'observa l'aparellament de les bases tipus Watson-Crick en el que seria la part estructurada central i les dues zones desestructurades en els extrems de l'oligonucleòtid on es produeix el gir de la cadena.

- Estructures triples

A l'any 1957 es van publicar els primers treballs que anunciaven la formació d'estructures triples en àcids nucleics [29, 30]. En els darrers anys ha ressorgit l'interès per l'estudi d'aquestes estructures degut a la seva possible implicació en processos biològics i a les seves potencials aplicacions biomèdiques com, per exemple, les teràpies antigèniques i antisentit [1].

Una hèlix triple apareix quan, per exemple, una cadena de nucleòtids es col·loca en el solc major d'una hèlix doble d'ADN i les seves bases interaccionen per ponts d'hidrogen tipus Hoogsteen amb les purines de l'aparellament Watson-Crick. L'estructura triple pot formar-se a través d'una única cadena polimèrica (estructura intramolecular) o a partir de diferents cadenes polimèriques (estructura intermolecular). La incorporació de la tercera cadena, també anomenada TFO (*Triplex Forming Oligonucleotides*), en el solc major de l'hèlix doble provoca un eixamplament d'aquest solc i la seva divisió en dos solcs asimètrics [1].

Les hèlixs triples en els àcids nucleics es divideixen en dos famílies, depenent de la identitat de les bases (purina o pirimidina) de la tercera cadena [1]. Així, com en els aparellaments de bases, tindrem les (pir·pur)∗pir on la tercera cadena és rica en pirimidines, i les (pir·pur)∗pur on la tercera cadena és rica en purines. A la família (pir·pur)∗pir, la tercera cadena és paral·lela a la cadena de purines i interacciona amb aquestes per ponts d'hidrogen Hoogsteen per donar les tríades (C·G)∗C^+ i (T·A)∗T (veure Figura 2.9.). La formació d'aquesta triada requereix la protonació del nitrogen 3 de la citosina de la tercera cadena i, per tant, presentarà un interval d'existència que dependrà del pH del medi. D'altra banda, la família (pir·pur)∗pur, la tercera cadena és antiparal·lela a

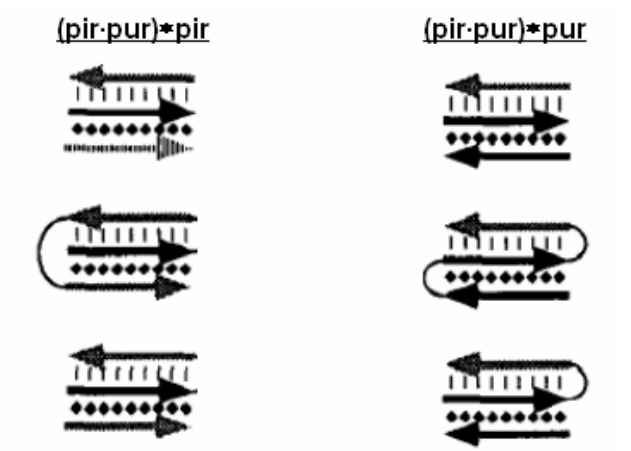


Figura 2.13. Representació de les estructures obtingudes en les famílies d'estructures triples. (|) representa un aparellament tipus Watson-Crick i (•) representa un altre tipus d'aparellament.

la cadena de purines del dúplex i interacciona amb aquesta per ponts d'hidrogen Hoogsteen invertit per formar les tríades (C·G)∗G, (T·A)∗A i (T·A)∗T. A la Figura 2.13. es mostren algunes de les opcions existents en la formació d'estructures triples.

- Estructures quàdruples

Hi ha diversos motius estructurals que poden adoptar les estructures formades per quatre cadenes.

G-quadruplex

En presència d'un ió monovalent les cadenes de nucleòtids riques en guanines poden agregar-se i formar una hèlix de quatre cadenes que es coneix com *G-quadruplex* [1, 9, 31, 32]. El *G-quadruplex* és una estructura inusual de l'ADN que es pot trobar a les regions telomèriques al final dels cromosomes. Durant el procés de replicació és necessari que existeixi una regió al final d'una de les cadenes d'ADN que serveixi com a encebador per a la síntesi de la cadena complementària. Aquestes regions extra són riques en guanines i la seva presència és molt important en el procés de replicació. El fragment extra amb un alt contingut de guanines es replega sobre sí mateix i les guanines s'aparellen mútuament per ponts d'hidrogen tipus Hoogsteen (Figura 2.10.) [1, 17]. Les quatre guanines, conegudes com *G-quartet*, es troben apilades unes sobre les altres a una distància de 3,4 Å.

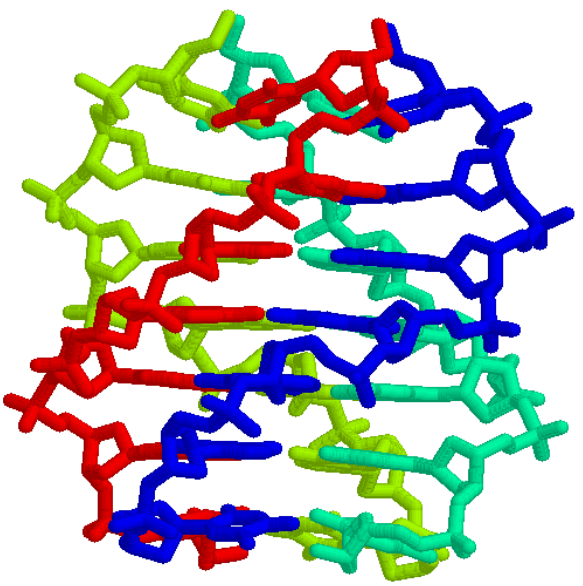


Figura 2.14. Estructura d'un *G-quadruplex*. Adaptat de l'arxiu PDB 139D [33] corresponent a [34] on cada color representa una cadena d'ADN.

Aquest tipus d'estructura pot formar-se a partir d'una única cadena (estructura intramolecular), o bé a partir de dues cadenes o quatre cadenes de nucleòtids (estructura intermolecular) [31, 32]. La disposició de les cadenes pot ser paral·lela o antiparal·lela depenent de la naturalesa de la seqüència o del medi iònic en el qual es trobi l'estructura. En el cas de la formació d'estructures quàdruples intermoleculars la posició relativa de les cadenes pot provocar la formació de dos tipus diferents d'estructures (veure Figura 2.15.). Així, hi hauran les estructures quàdruples paral·leles en les quals les quatre cadenes es trobaran en el mateix sentit i les estructures quàdruples antiparal·leles en les quals les cadenes es trobaran en sentits diferents dos a dos. Les estructures quàdruples antiparal·leles són les més àmpliament estudiades a la literatura degut a que són les que formen les estructures quàdruples intramoleculars [32].

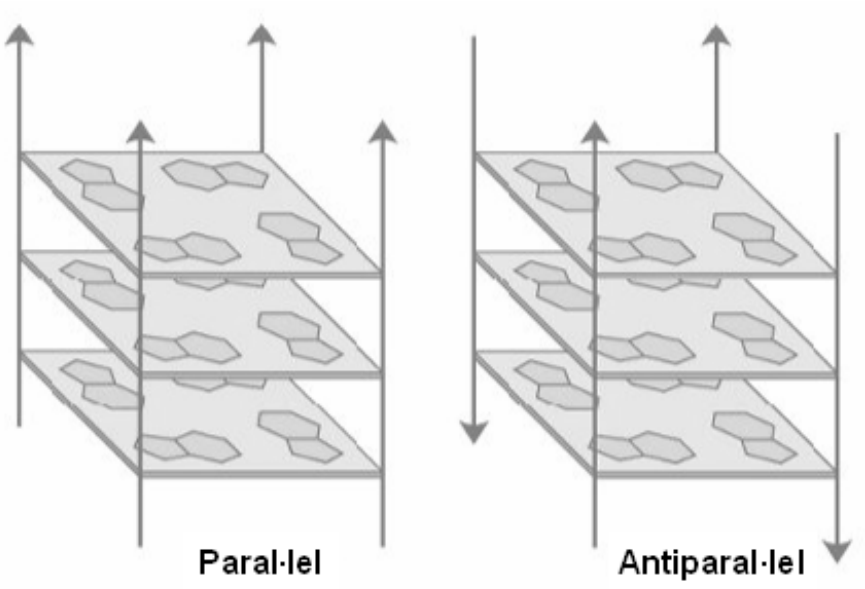


Figura 2.15. Esquema d'estructures quàdruples en disposicions paral·leles i antiparal·leles.

Bi-loop

A diferència de les estructures *G-quadruplex*, la formació d'estructures tipus *bi-loop* no necessita cadenes amb un contingut elevat en guanina sinó que es formen per aparellaments de bases tipus Watson-Crick i interaccions degudes a l'apilament de bases [35, 36]. A la bibliografia es troben descrites com dímers d'estructures cícliques o lineals en les quals es formen quatre parells de bases de tipus Watson-Crick intermoleculars que permeten l'estabilitat de l'estructura formada per les quatre cadenes. Malgrat no tenir una implicació biològica tan obvia com en el cas dels *G-quadruplex*, aquest motiu estructural pot estar implicat en processos de reconeixement

de molècules d'àcids nucleics com, per exemple, la reordenació genètica que es produeix durant la formació de l'esperma [36].

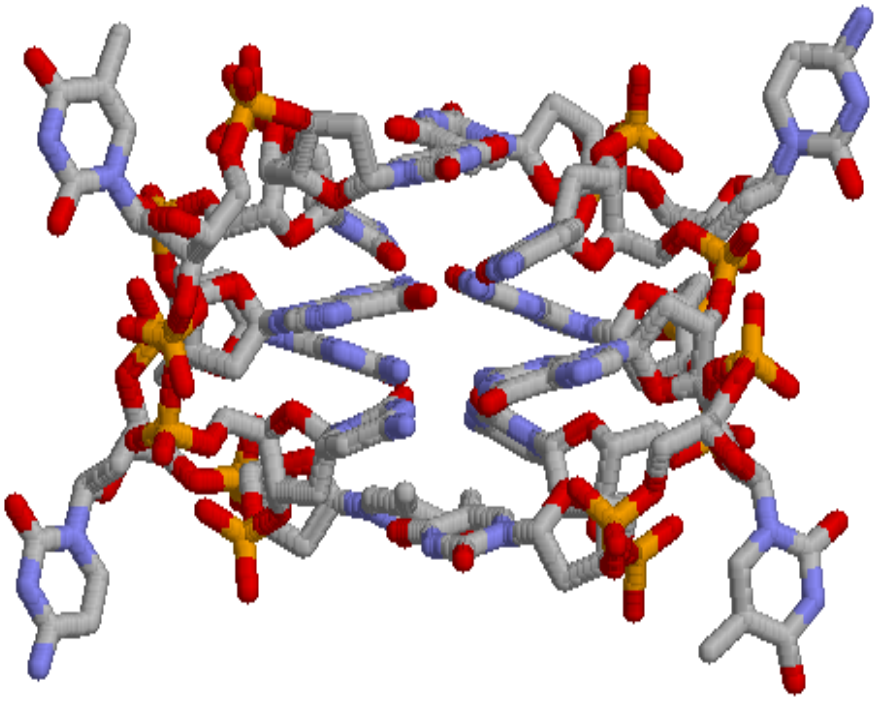


Figura 2.16. Estructura quàdruple tipus *bi-loop* adaptada de [37]

A la Figura 2.16. s'observa un exemple d'una estructura bi-loop. Degut a l'apantallament de les repulsions dels grups fosfat degut als efectes dels cations metàl·lics, les dues cadenes de l'estructura tipus *dumbbell* es poden apropar fins que es formen ponts d'hidrogen intermoleculars adoptant una estructura quàdruple.

2.2. Variables experimentals que afecten a les conformacions dels àcids nucleics

Com hem vist a l'apartat anterior les diferents conformacions que poden adoptar els àcids nucleics es basen en els parells de bases estabilitzats per ponts d'hidrogen. Donada la feblesa d'aquest tipus d'enllaços és lògic pensar que canvis en el medi podran afectar aquests enllaços i, per tant, provocar canvis en l'estructura adoptada. Així, entre els factors experimentals que poden afectar la conformació de l'ADN destaquen els següents: temperatura, pH, força iònica (tant per la concentració de sals

com pel tipus de sal), concentració de l'àcid nucleic, polaritat del solvent o presència de ions metàl·lics [16, 17].

Al ser els àcids nucleics polímers, s'han de tenir en compte dos efectes relacionats amb la complexitat d'aquests [16]. En primer lloc, els efectes polifuncionals que són deguts a la presència de grups funcionals veïns amb propietats químiques diferents. Així, grups carboxil, amida, amino, grups fosfat i grups hidroxil lliures als extrems de les cadenes. Tota aquesta varietat de grups funcionals diferents implica que, al estudiar un procés, es pot estar observant una suma de diferents processos simultanis. Per exemple, la protonació o desprotonació de les bases nitrogenades pot possibilitar o impossibilitar la formació de qualsevol dels parells de bases descrits anteriorment. En segon lloc, i molt relacionat amb l'anterior, s'hi troben els efectes polielectrolítics, que es produeixen per l'aparició o desaparició de càrregues elèctriques en els grups funcionals del polímer. Un canvi de pH, un procés de complexació o una modificació de la força iònica del medi pot provocar canvis en la compensació parcial de les càrregues iòniques que apareixen al polímer induint canvis conformacionals degut a l'atracció/repulsió entre les càrregues.

Donada la complexitat estructural dels àcids nucleics, els efectes secundaris que afecten un grup funcional i/o un parell de bases es troben àmpliament influïts per la presència de grups funcionals pròxims, l'apilament de bases nitrogenades i de les interaccions per ponts d'hidrogen existents. En aquest casos, els efectes descrits anteriorment produeixen un comportament cooperatiu a la molècula. L'existència d'aquests processos cooperatius provoca que els canvis en les estructures dels àcids nucleics siguin bruscos, degut a que la formació o el trencament dels enllaços entre un parell de bases indueix a la formació o al trencament dels enllaços propers. La importància d'aquests efectes és proporcional a la longitud de la cadena dels àcids nucleics i, per tant, de la complexitat estructural dels àcids nucleics. Així, aquests efectes tindran gran importància en els polinucleòtids però en els oligonucleòtids (amb els quals s'ha treballat majoritàriament en aquesta Tesi Doctoral) aquests efectes poden arribar a ser pràcticament menyspreables.

A continuació es descriuen els principals factors experimentals que poden afectar l'estabilitat de les conformacions dels àcids nucleics.

Efecte de la temperatura

Les conformacions dels àcids nucleics presenten una gran dependència amb la temperatura del medi [17]. Així, a temperatures baixes els àcids nucleics acostumen a trobar-se en una conformació estructurada mentre que a temperatures més elevades es troben en una conformació desestructurada. Aquesta transició des d'una conformació estructurada a una conformació desestructurada mitjançant l'augment de la temperatura del sistema és el que es coneix com desnaturalització tèrmica (*melting*). En el cas dels àcids nucleics es pot dur a terme el procés contrari, conegut com renaturalització (*annealing*), en el qual en baixar la temperatura del medi es produeix la transició de conformació desestructurada a conformació estructurada. El procés de desnaturalització-renaturalització acostuma a ser reversible ja que la velocitat a la qual s'escalfa o es refreda la solució és més lenta que la velocitat a la qual es produeix l'equilibri entre les conformacions de l'àcid nucleic presents en el sistema [9, 38].

Aquesta transició entre les conformacions estructurada i desestructurada es produeix en un interval estret de temperatures (Figura 2.17.), i permet determinar el que es coneix com temperatura de fusió o *melting* (T_m) que correspon a la temperatura a la qual la meitat de l'àcid nucleic es troba en la conformació estructurada i l'altre meitat en la conformació desestructurada.

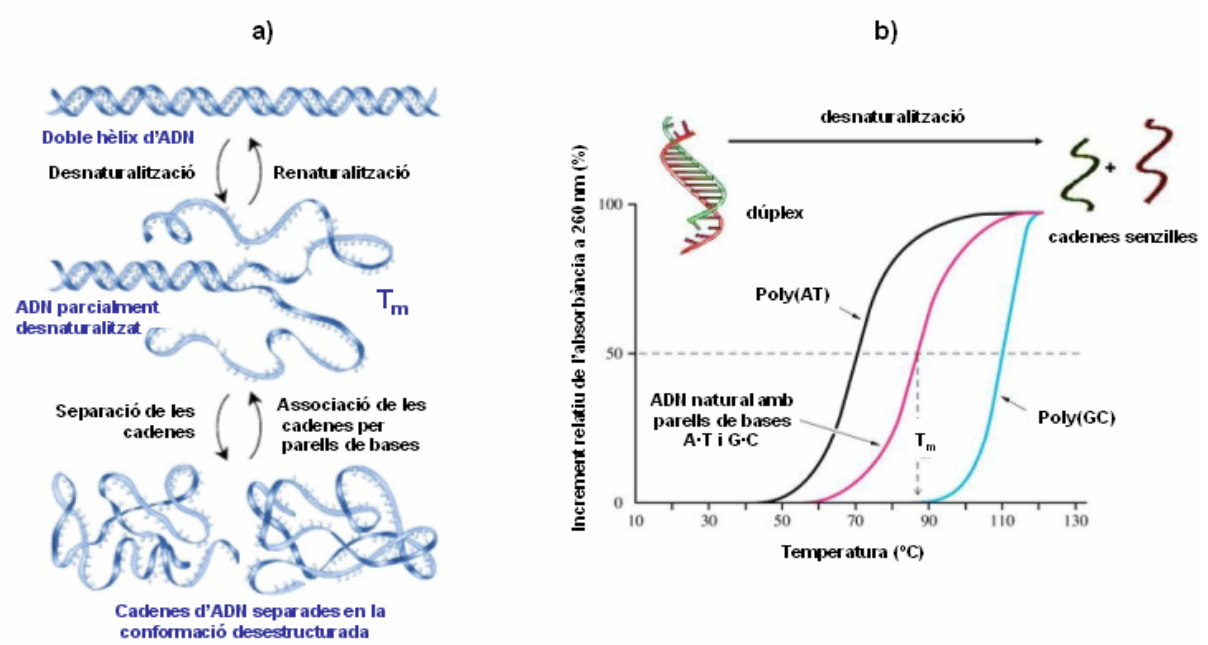


Figura 2.17. a) Representació esquemàtica d'un experiment de desnaturalització i, b) representació de la corba de desnaturalització i determinació del valor de T_m adaptat de [39].

El procés de desnaturalització (o renaturalització) d'un àcid nucleic acostuma a provocar canvis en les propietats espectrals. Aquests canvis poden ser seguits mitjançant diferents tècniques; com ara l'absorció molecular a l'ultraviolat (on generalment es treballa a 260 nm), la fluorescència, la ressonància magnètica nuclear, el dicroisme circular o l'espectroscòpia Raman [17].

Habitualment si es segueix la transició entre les conformacions estructurades i desestructurades mitjançant absorció molecular s'observa el fenomen de la hipercromicitat [40]. Aquest és l'augment de l'absorció de les bases de l'àcid nucleic en augmentar la temperatura (Figura 2.18.a). Aquest fet és degut a l'afebliment de les interaccions per apilament entre les bases nitrogenades veïnes. Cal destacar que aquesta hipercromicitat es troba, en general, al realitzar experiments de desnaturalització d'estructures dúplex a cadenes desestructurades. Si s'estudien altres estructures la hipercromicitat pot ser mínima o, fins i tot, es pot produir-se el fenomen contrari conegut com a hipocromicitat. Una altra tècnica molt utilitzada per estudiar aquests canvis conformacionals és el dicroisme circular [41]. En aquest cas, també, es sol observar una disminució de la intensitat del senyal en augmentar la temperatura però acompanyada de canvis significatius en la forma de l'espectre obtingut (Figura 2.18.b). Aquests fets es produeixen per la creixent desorganització de la molècula que provoca una menor simetria en la mateixa i, en conseqüència, un senyal de dicroisme circular diferent.

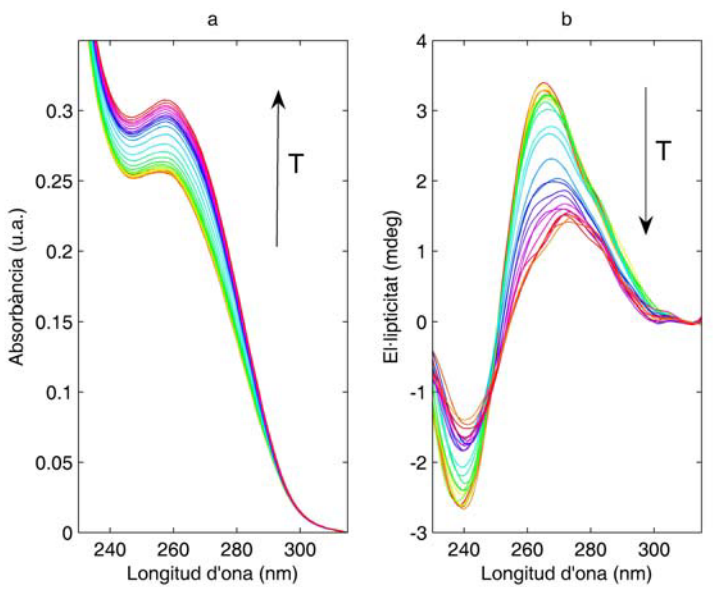


Figura 2.18. Espectres obtinguts en augmentar la temperatura. a) Augment de l'absorbància a l'absorció molecular i , b) Disminució del senyal de dicroisme circular.

El valor de la temperatura de fusió (T_m) que es determina té un interès limitat en el camp bioquímic en el qual habitualment es treballa a la temperatura fisiològica. Per tal d'extreure més informació, s'han establert les equacions per determinar paràmetres termodinàmics com l'entalpia, l'entropia estàndard i l'energia lliure de Gibbs a partir de la corba de fusió [38, 42]. Aquests càlculs es basen en l'establiment d'un model de dos estats (totalment estructurat i totalment desestructurat) la qual cosa pot provocar l'aparició d'errors importants si existeixen intermedis en concentracions que no siguin menyspreables. També és necessari un coneixement previ del sistema ja que les equacions a utilitzar seran diferents depenent de la molecularitat i l'estequiometria de la reacció. Per exemple, en el cas més senzill de dues úniques conformacions es pot definir la fracció desestructurada com una funció de la hipercromicitat (és a dir, de l'absorbància (A), veure corba a la Figura 2.17.b):

$$f = \frac{A_{experimental} - A_{estructurat}}{A_{desnaturalitzat} - A_{estructurat}} \qquad \text{Equació 2.1.}$$

A partir d'aquí es pot determinar el valor de la constant d'equilibri en cada punt. Així per a la dissociació d'una estructura de doble hèlix es tindrà:

$$K(T) = \frac{[Desestructurat]^2}{[Estructurat]} = \frac{(1-f)^2}{2f} c_T \qquad \text{Equació 2.2.}$$

On a partir de la fracció desestructurada (f) i de la concentració total d'oligonucleòtid (c_T) es pot obtenir el valor de la constant d'equilibri a cada temperatura. L'Equació 2.2. pot adoptar diferents formes en funció de les característiques del sistema estudiat. Per exemple, en el cas d'oligonucleòtids autocomplementaris seria:

$$K(T) = \frac{[Desestructurat]^2}{[Estructurat]} = \frac{2(1-f)^2}{2f} c_T \qquad \text{Equació 2.3.}$$

A partir de l'estudi de la dependència de K amb la temperatura es poden obtenir els paràmetres termodinàmics utilitzant l'equació clàssica:

$$\Delta G^{o} = -RT\ln(K(T)) = \Delta H^{o} - T\Delta S^{o}$$ Equació 2.4.

A la pràctica, l'aplicació d'aquesta fórmula es fa més fàcil en representar la constant d'equilibri amb l'invers de la temperatura. Així, es podran extreure els paràmetres termodinàmics a partir de l'ordenada a l'origen i la pendent de la recta obtinguda.

$$\ln(K(T)) = \frac{\left(\frac{-\Delta H^{o}}{R}\right)}{T} + \frac{\Delta S^{o}}{R}$$ Equació 2.5.

Efecte de la concentració salina

Les bases nitrogenades són neutres en el rang de pH fisiològic (entre 5 i 9). El mateix passa amb les pentoses, que només poden perdre el protó del grups hidroxils a valors de pH extremadament bàsics. En canvi, el grup fosfat presenta una càrrega negativa, per la qual cosa els àcids nucleics es poden considerar com a polianions que interaccionen amb molècules carregades positivament com, per exemple, els cations. Aquesta interacció electrostàtica apantalla les repulsions entre les càrregues negatives dels grups fosfat estabilitzant les estructures ordenades de l'ADN [1].

Entre les possibles interaccions entre la cadena d'ADN i els cations cal destacar la condensació de contraions [17]. En aquesta, els cations es troben a molt poca distància de l'ADN però amb llibertat per moure's de forma paral·lela a l'hèlix doble. A més d'aquesta interacció deslocalitzada, alguns ions tenen preferència per interaccionar de manera localitzada amb determinats grups funcionals. Així, per exemple, els cations més habituals en els estudis d'equilibris entre conformacions (Na^{+}, K^{+}, Mg^{2+}, Ca^{2+}) tenen preferència per enllaçar-se als grups fosfats.

En general, l'efecte estabilitzador dels cations divalents és més gran que l'efecte dels cations monovalents. Aquest fet es pot explicar perquè la capacitat apantalladora dels ions divalents és molt més gran que la dels ions monovalents. Malgrat això, en alguns casos com els *G-quadruplex* l'efecte estabilitzador és més gran en el cas del cations monovalents perquè, a més de la càrrega, s'ha de tenir en compte el tamany del ió. Així, depenent de l'estructura formada i de la seqüència, es podran observar preferències per ions amb un radi més gran o més petit.

Efecte del pH

Un altre factor que afecta les conformacions dels àcids nucleics és el pH de la solució [17]. Per això, molts experiments que es realitzen al laboratori utilitzen dissolucions amortidores que permeten mantenir el pH de la solució al voltant d'aquests valors de pH. En la literatura es poden trobar diferents exemples d'aquests dissolucions amortidores entre els quals destaquen el tampó de fosfats, el PIPES (piperazina-N,N'-bis[àcid etanosulfònic]) o el cacodilat de sodi.

Per una altra banda, és interessant conèixer si es produeixen canvis conformacionals al variar el pH del medi. Al modificar el pH de la solució existeix la possibilitat de que es produeixi la protonació/desprotonació dels àtoms de nitrogen que es troben a les bases (Taula 2.1.) [5]. Quan es produeix aquesta protonació/desprotonació, l'estabilitat del parell de bases corresponent es veurà afectada i això pot implicar canvis conformacionals. Un clar exemple dels efectes de la modificació de pH és la formació d'hèlixs triples, amb la triada de bases $(C{\cdot}G){*}C^{+}$, a partir d'una hèlix doble i que necessita la protonació de les citosines de l'altre cadena [1].

A la taula següent es mostres els valors de pK_a per a la ionització de les bases nitrogenades.

Taula 2.1. Taula de valors de pK_a dels fosfats i de les bases nitrogenades dels àcids nucleics adaptada de [5].

	Fosfat		**Base**	
	Ionització Primària	Ionització Secundària	pK_a	N de la base on es perd el protó
	pK_{a1}	pK_{a2}		
5' AMP	0,9	6,1	3,8	N1
5' GMP	0,7	6,1	2,4	N7
			9,4	N1
5' TMP	1,0	6,4	9,5	N3
5' CMP	0,8	6,3	4,5	N3

Efecte de la concentració dels àcids nucleics

La concentració d'àcid nucleic en la solució és un altre factor a tenir en compte. Habitualment es treballa amb concentracions d'àcid nucleics de l'ordre micromolar ja que, a partir d'aquesta concentració, les senyals espectroscòpiques obtingudes acostumen a ser bones. Per una altra banda, el treball amb concentracions més elevades pot donar lloc a fenòmens d'agregació [2]. Així, si s'augmenta un o dos ordres de magnitud la concentració de treball s'augmenten, al mateix temps, les possibilitats de formació d'estructures intermoleculars d'ordre superior, és a dir, la formació d'estructures tricatenàries o tetracatenàries.

Es pot conèixer la formació d'estructures intermoleculars mitjançant la realització d'experiments de desnaturalització tèrmica a diferents concentracions d'àcid nucleic [17, 42]. Per exemple, la desnaturalització d'una estructura unimolecular no depèn de la concentració i, per tant, no observarem canvis en la temperatura de fusió en augmentar la concentració d'àcid nucleic. En canvi, la fusió d'una estructura bimolecular depèn de la concentració d'àcid nucleic. En aquest cas, en augmentar la concentració d'àcid nucleic s'observa que la temperatura de fusió augmenta. Això es deu a que en tenir més molècules monomèriques en solució, la probabilitat de formació de ponts d'hidrogen intermoleculars també augmenta i, per tant, l'enllaç entre dues molècules es pot mantenir fins a temperatures més elevades. Aquest fet es pot aprofitar per determinar els paràmetres termodinàmics associats a un procés multimolecular. L'equació per calcular ΔH^o i ΔS^o, en aquest cas, és:

$$\frac{1}{T_m} = \frac{R\ln(c)}{\Delta H^o} + \frac{\Delta S^o}{\Delta H^o} \qquad \text{Equació 2.6.}$$

2.3. Descripció dels sistemes experimentals estudiats

Tres dels treballs inclosos en el capítol 5 d'aquesta memòria mostren els resultats obtinguts en l'estudi dels equilibris conformacionals induïts pels canvis de temperatura, pH i força iònica del medi per diferents oligonucleòtids. Així, s'ha realitzat l'estudi experimental dels sistemes següents:

Estructures dumbbell* / *bi-loop

En el primer treball (apartat 5.2.) s'estudien els equilibris en solució de l'oligonucleòtid cíclic d<pTGCTCGCT>. Aquest oligonucleòtid adopta una estructura tipus *dumbbell* en medi aquós a força iònica i temperatura baixes. A més, l'oligonucleòtid estudiat adopta una altra estructura en medi aquós a baixes temperatures però a forces iòniques elevades [37, 43]. En aquestes condicions es forma un dímer del tipus *bi-loop d*egut a l'apantallament de les repulsions dels grups fosfat pels cations metàl·lics. D'aquesta forma les dues cadenes de l'estructura *dumbbell* es poden apropar més fins que es formen ponts d'hidrogen intermoleculars i es forma una estructura quàdruple intermolecular.

L'objectiu principal d'aquest treball era estudiar els equilibris entre aquestes dues estructures. Així, es volia estudiar la seva estabilitat tèrmica en condicions tant de força iònica baixa, on s'esperava trobar l'estructura tipus *dumbbell,* com en condicions de força iònica elevada, on s'esperava trobar l'estructura tipus *bi-loop*.

Estructures triples

En el segon treball (apartat 5.3.) s'estudien els equilibris àcid-base de quatre oligonucleòtids i les possibles interaccions entre ells. Es va treballar amb la seqüència h26 (5'–GAAGGAGGAGA–TTTT–TCTCCTCCTTC–3') i les seqüències s11AG (5'–AGAGGAGGAAG–3'), s11TG (5'–TGTGGTGGTTG–3') i s11CT (5' – CTTCCTCCTCT – 3').

La seqüència h26 presenta una primera estructura de forqueta amb el gir format per les quatre bases de timina i una segona estructura d'hèlix doble (veure, per exemple la Figura 2.11.). En canvi, els altres tres oligonucleòtids no presenten una estructura secundària definida i es troben en solució en forma d'una única cadena. La interacció de cadascuna d'aquestes tres cadenes d'oligonucleòtid amb l'estructura en forma de forqueta permet l'obtenció de dos tipus diferents d'estructures triples (paral·lela i antiparal·lela) que presenten unes característiques biofísiques totalment diferents [1, 44, 45].

L'objectiu principal d'aquest treball va ser estudiat la dependència de la formació de les estructures triples amb el pH. A més es pretenia estudiar l'estabilitat tèrmica de cadascuna d'aquestes estructures triples formades per tal de poder avaluar la seva possible importància biomèdica.

Estructures quàdruples

Finalment, a l'apartat 5.4., es van estudiar els canvis conformacionals induïts per la temperatura de l'oligonucleòtid 5'–TAGGGTTAGGGT–3'. S'ha observat que la interacció de dues molècules a temperatura baixa permet formar una estructura de *G-quadruplex* intermolecular en la qual es formen aparellaments quàdruples de bases de guanina enllaçades per ponts d'hidrogen [32, 46]. A la literatura s'ha descrit que poden existir estructures diferents simultàniament en solució depenent de la temperatura i de la presència de cations monovalents.

L'objectiu principal d'aquest treball ha estat la resolució de totes les estructures presents en diferents condicions experimentals.

2.4. Micromatrius d'ADN

2.4.1. Conceptes bàsics

El 24 d'abril de l'any 2003 es va publicar la seqüenciació dels aproximadament 30000 gens que formen el genoma humà gràcies al programa *Human Genome Project* (HGP) [47-49]. Encara que aquest projecte va ser iniciat als anys vuitanta sota la direcció de James Watson, va ser a partir de la dècada dels noranta quan l'aparició de noves tècniques instrumentals (com la reacció en cadena de la polimerasa [50, 51], PCR) van permetre un ràpid progrés, que va culminar amb la seqüenciació completa del genoma humà. El següent pas consisteix en localitzar i determinar la funció dels gens [52]. Donada la necessitat d'analitzar un gran nombre de dades, han aparegut noves tècniques instrumentals que permeten aquests estudis. Entre aquestes tècniques, destaca la tecnologia de les micromatrius d'ADN [53, 54]. La tecnologia de les

micromatrius d'ADN permet l'estudi massiu i simultani de molts gens, tot proporcionant l'oportunitat d'estudiar la seva expressió gènica i les possibles interaccions entre ells.

L'aplicació més popular de la tecnologia de les micromatrius d'ADN és el conegut com *gene discovery* [55] que permet la identificació de nous gens i ampliar el coneixement sobre el seu funcionament i nivells d'expressió en condicions biològiques diferents. Per exemple, en estudiar com és l'expressió d'un grup de gens en teixits normals i cancerosos es pot observar que l'expressió d'alguns gens és significativament diferent en els dos teixits, per la qual cosa aquest gens es consideraran sospitosos d'estar relacionats amb l'aparició del càncer. Una altra aplicació és el diagnòstic de malalties [52, 56, 57]. En aquest cas, l'anàlisi dels experiments de micromatrius permet identificar els gens relacionats amb malalties provocades pel medi ambient com, per exemple, malalties que tenen a veure amb els sistemes respiratori, nerviós o immunològic. Múltiples treballs s'han publicat estudiant diferents tipus de càncer amb base a l'expressió gènica en les cèl·lules tumorals. Una altra aplicació és el desenvolupament de fàrmacs [10, 52, 58, 59] (*farmacogenomics*) que estudia la possible relació existent entre l'administració d'un fàrmac i els perfils genètics del pacient. Una anàlisi comparativa dels gens d'una cèl·lula malalta i una cèl·lula normal ajuda a identificar la constitució bioquímica de les proteïnes sintetitzades a partir de gens de la cèl·lula malalta. Els investigadors poden utilitzar aquesta informació per sintetitzar fàrmacs amb els que combatre aquestes proteïnes i reduir els seus efectes. Finalment, la investigació toxicològica [60, 61] (*toxicogenomics*) que estudia les correlacions entre les respostes a substàncies tòxiques i els canvis en els perfils genètics dels organismes exposats a aquestes. La tecnologia de les micromatrius d'ADN facilita els tests de toxicitat dels fàrmacs, de forma que la determinació d'efectes col·laterals i incompatibilitats potencials dels fàrmacs en una etapa inicial pot ajudar a estalviar temps i diners degut a la selecció de fàrmacs amb més possibilitats d'èxit.

Els orígens de les micromatrius d'ADN s'han de buscar a la dècada dels setanta quan Ed Southern va començar a realitzar estudis d'hibridació entre seqüències d'ADN lliures i seqüències unides a un suport sòlid [62]. No va ser fins l'any 1995 quan el grup del Dr. Pat O. Brown a la Universitat de Stanford [63] va descriure els *microarrays* d'ADN tal i com es coneixen actualment.

Bàsicament, una micromatriu és una placa suport revestida de vidre o plàstic en la qual les seqüències d'ADN es fixen a determinats punts anomenats pouets (*spots*) (Figura 2.19.) [53]. En cadascuna de les plaques poden ser impresos fins a 20000 *spots* de forma que es podran analitzar fins a 20000 seqüències, ja siguin de gens coneguts o del que es coneix com *Expressed Sequence Tags* (EST) que són seqüències d'ADN que s'expressen però que se'n desconeix la funció.

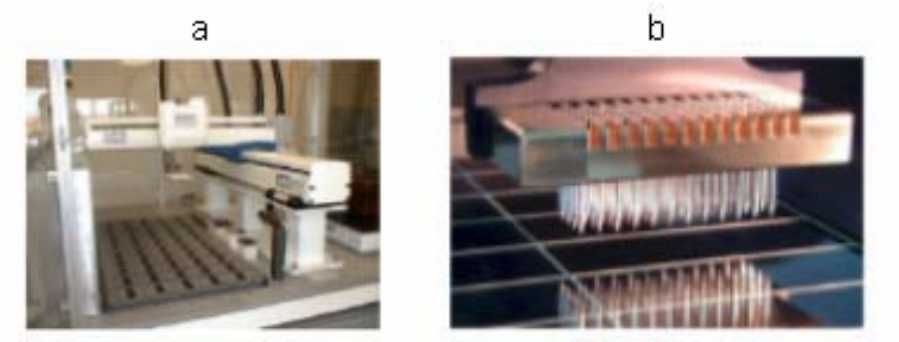


Figura 2.19. Instrumentació utilitzada en un experiment de microarrays d'ADN adaptada de [64]. a) Instrument de deposició de les sondes a la placa (*spotting*) i b) exemple de capçal d'impressió de les sondes.

Hi ha dos tipus bàsics d'experiments de micromatrius d'ADN:

1) Sondes d'ADN complementari, de entre 500-5000 bases, i que van ser desenvolupades a la Stanford University.

2) Graelles d'oligonucleòtids (o xips d'ADN), de entre 20 i 100 bases, que van ser desenvolupades per la casa comercial Affymetrix, Inc. que ven el producte sota el nom de GeneChip®.

A continuació es farà una breu descripció de procés experimental necessari en el cas de dur a terme un experiment de micromatrius d'ADN complementari seguint l'esquema que es mostra a la Figura 2.20. [65].

En primer lloc, es preparen les sondes o clons d'ADN complementari. Per això es defineixen i es seleccionen els gens que s'estudiaran a l'experiment. Per dur a terme aquesta selecció es poden fer servir les bases de dades aparegudes arrel del projecte de seqüenciació del genoma humà. En aquestes bases de dades hi ha milers de

seqüències conegudes de bases d'ADN les quals corresponen a fragments d'ADN que codifiquen un gen o una etiqueta d'expressió de la seqüència (EST). Aquestes etiquetes d'expressió de la seqüència corresponen a seqüències de bases d'ADN de mida variable i que, en alguns casos, s'ha vist posteriorment que corresponen a gens presents en el genoma humà [65, 66]. El següent pas és la immobilització d'aquestes sondes o clons en el suport sòlid utilitzat que acostuma a ser de vidre. En cada pouet del suport es dipositen fraccions de nanolitres de les sondes prèviament obtingudes, purificades i ampliades per PCR. Aquestes sondes es distribueixen en plaques d'aproximadament 400 pouets, és a dir, un experiment complet de micromatrius d'ADN generalment farà servir més d'una placa.

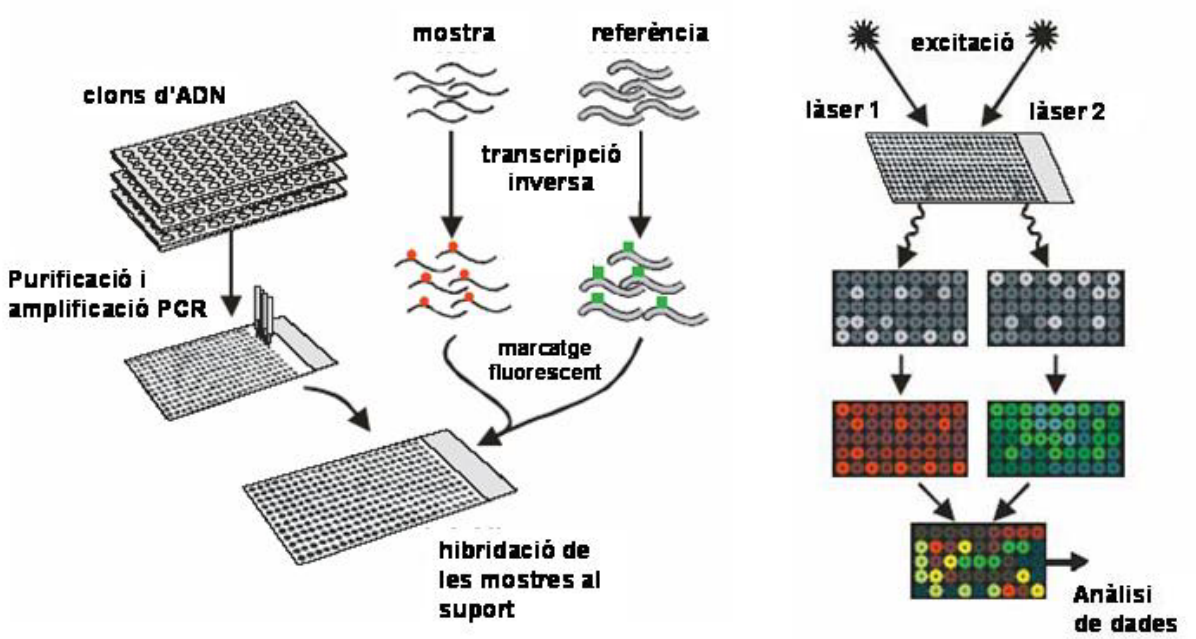


Figura 2.20. Esquema general del procés de fabricació i mesura d'una micromatriu d'ADN del tipus sondes d'ADN complementari adaptada de [65].

D'altra banda, es preparen els ADN complementaris de mostres problema i de referència. En aquest procés es veu reflectit el Dogma central de la Biologia Molecular que s'ha mostrat a la Figura 2.1. ja que a partir de l'ADN es prepara l'ARN missatger i d'aquest s'obté el seu ADN complementari. La generació d'ADN complementari a partir d'ARN missatger es porta a terme mitjançant transcripció inversa amb la peculiaritat de que es marca l'ADN complementari amb fluoròfors. Una vegada sintetitzats els ADN complementaris problema i referència es marquen, respectivament, un amb el fluoròfor verd Cye-3 i l'altre amb el fluoròfor vermell Cye-5 [65, 66]. Generalment es fa servir el

colorant vermell per a l'ADN complementari de la mostra problema i el colorant verd per a l'ADN complementari de referència.

El següent pas és dur a terme la hibridació, és a dir, la unió mitjançant ponts d'hidrogen del l'ADN complementari de les mostres problema i de referència amb les seqüències sonda d'ADN prèviament unides al suport. Així, les sondes d'ADN presents a cada pouet del suport competeixen pels ADN complementaris marcats de les mostres problema i de referència. Durant aquest procés es poden produir quatre situacions diferents a cada pouet. En primer lloc, hi ha el cas que no es produeixi hibridació ni de la mostra problema ni de referència. En segon lloc, pot donar-se que únicament s'hibridi una de les mostres, és a dir, que s'hibridi per ponts d'hidrogen o la cadena de la mostra problema o la cadena de la mostra de referència. Finalment, hi ha el cas on a un determinat pouet s'hibriden les dues cadenes (tant les de mostra com la de referència), malgrat que pot haver-hi preferència per una de les dues. El darrer pas, abans del procés de mesura espectroscòpica, és el rentat de les restes d'ADN complementari que no s'han hibridat específicament [53].

L'excitació làser de la micromatriu proporciona una emissió fluorescent amb uns espectres característics els quals són mesurats utilitzant un microscopi d'escaneig làser confocal. Les imatges monocromes (una corresponent al colorant verd i l'altre al colorant vermell) del escàner són importades pel programari en el qual són pseudocolorejades i fusionades. Així, d'acord amb les quatre possibilitats que s'han descrit al paràgraf anterior s'obtindrà una expressió gènica determinada pel gen o EST que es troba en un determinat pouet. D'aquesta forma, un pouet apareixerà de color negre si no s'han hibridat ni l'ADN complementari problema ni l'ADN complementari de referència. El pouet apareixerà de color verd si només s'ha hibridat l'ADN complementari de la mostra problema i vermell si només s'ha hibridat l'ADN complementari de la mostra de referència. En el cas, que s'hidridin les mostres problema i de referència al mateix pouet s'obtindrà un senyal de color groc malgrat que depenent de l'abundància pot existir una tonalitat vermella o verda. D'aquesta forma, els nivells de l'expressió gènica relativa a les mostres problema i a les mostres de referència poden ser estimades a partir de les intensitats de fluorescència i els colors emesos per cada pouet durant l'escaneig [54, 65].

Aquestes imatges escanejades són analitzades utilitzant programari d'anàlisi d'imatges. Aquests programes avaluen l'expressió dels gens quantificant la relació de la intensitat de fluorescència per a cada pouet. Així, les intensitats quantificades proporcionen informació sobre l'activitat d'un gen específic en una determinada cèl·lula o teixit. A partir d'aquestes relacions d'intensitats de fluorescència es realitza l'anàlisi de les dades experimentals per tal de determinar les possibles correlacions entre les expressions dels gens.

2.4.2. Tractament de dades

L'expressió relativa dels gens en una sèrie de mostres es pot tabular en forma de matriu de dades [10]. Així, la matriu de dades **D**, les dimensions de la qual serà de *n* gens per *m* mostres en condicions diferents, representarà la totalitat de les dades d'expressió on cada valor serà el $\log_2$ de la relació de les expressions.

$$x_{ij} = \log_2 \frac{C5_{ij,\,vermell}}{C3_{ij,\,verd}} \qquad \text{Equació 2.7.}$$

on $C5_{ij}$ és la intensitat de fluorescència del gen *i* a la mostra problema *j* i $C3_{ij}$ és la intensitat de fluorescència del gen *i* a la mostra de referència *j*.

Així, x_{ij} serà negatiu si C3>C5 i positiu si C5>C3, el que equival a dir que un gen estarà sobreexpressat respecte la mostra de referència quan x_{ij} sigui positiu i un gen estarà subexpressat respecte la mostra de referència quan x_{ij} sigui negatiu.

Fonts de variabilitat i pretractaments de les dades

En els experiments de micromatrius d'ADN es tracta de determinar si la variació que es pot observar en les dades és deguda a variacions en els processos biològics entre la mostra problema i la mostra de referència [10]. En els primers treballs que van aparèixer utilitzant la tecnologia de les micromatrius d'ADN únicament es publicaven les relacions entre les intensitats de forma que no es podia conèixer la qualitat de les dades originals. Actualment aquest fet està canviant degut a la implantació del protocol MIAME (*Minimum Information About a Microarray Experiment*) [67, 68] que defineix els

estàndards mínims per publicar dades de micromatrius d'ADN. D'aquesta forma, un dels principals objectius en el disseny d'experiments de micromatrius d'ADN serà intentar assegurar que les fonts de variació no relacionades amb els processos biològics tinguin un efecte mínim. Entre aquestes fonts de variabilitat es pot destacar [10, 53]:

- L'extracció de l'ARN dels teixits, on factors com la quantitat de mostra o la puresa de l'ARN extret poden afectar a les posteriors etapes de mesura.
- El procés d'hibridació, que es pot veure afectat per diferents variables experimentals com la temperatura o el temps, els quals poden provocar resultats diferents en experiments replicats.
- El marcatge de les sondes. Així, es poden tenir diferències de resposta entre els pouets d'una mateixa placa ja que la seva situació no és exactament la mateixa i entre els pouets de plaques diferents.
- El procés d'escaneig pot tenir fonts de variació relacionades amb la instrumentació a causa de les possibles diferències en els rendiments dels làsers i dels detectors.
- L'anàlisi de les imatges obtingudes, on s'ha de quantificar cada pouet a partir dels píxels que el formen. Es necessiten diverses mesures per a cada punt, a més d'una mesura del fons per tal d'obtenir la intensitat neta.

Per tal d'aconseguir minimitzar aquestes fonts de variabilitat el disseny d'un experiment de micromatrius d'ADN inclou la mesura de rèpliques. Aquestes rèpliques poden ser de dos tipus diferents. En primer lloc, els replicats tècnics amb els quals es controla l'error experimental. En segon lloc, els replicats biològics, en els quals dos o més teixits que provenen de la mateixa font i que es tracten de forma idèntica serveixen per determinar la variabilitat global de les dades.

Per tal de minimitzar les variacions implícites a l'experiment que poden ser tant sistemàtiques com aleatòries es realitzen pretractaments de les dades com a pas previ de l'anàlisi estadístic [10, 69]. Així, s'acostuma a centrar les dades respecte la mitjana o la mediana de cada fila de la matriu de dades que les conté, la qual cosa implica que la mitjana o mediana de cada fila serà igual a zero. Per exemple, es troben a la bibliografia molts treballs que consisteixen en una sèrie de mostres de teixits cancerosos que es comparen amb una mostra de referència de teixit no cancerós [70-72]. També existeix la possibilitat de centrar les dades respecte a les columnes per

eliminar alguns tipus de biaix que desplaci tots els valors de la relació de fluorescències en un valor determinat. En segon lloc, es poden dividir tots els elements d'una fila de la matriu de dades de forma que la seva desviació estàndard sigui igual a la unitat. Aquests diferents tipus de pretractaments de les dades afecten la variabilitat de les dades i permeten amplificar senyals dèbils o disminuir senyals forts [10, 54]. Alternativament, existeixen altres mètodes de pretractament com poden ser els de suavitzat LOWESS (LOcally WEghted Scatterplot Smoothing) o la normalització respecte la intensitat total.

2.4.2.1. Mètodes de classificació

Actualment, un dels objectius que s'intenta assolir quan s'analitzen dades de micromatrius d'ADN és la classificació de les mostres estudiades segons aspectes biològics. Així, s'intenta agrupar les mostres en un nombre petit de grups que permeti disminuir la complexitat de les dades i aconseguir extreure similituds o diferències entre mostres d'una forma més intuïtiva. A continuació, es descriuen breument alguns dels mètodes més utilitzats per trobar aquests tipus d'agrupaments, els quals formen part del que s'anomena anàlisi d'agrupacions (*Cluster Analysis*). L'anàlisi d'agrupacions es pot definir com el procés de classificació d'una sèrie d'objectes en grups basant-se en relacions de similitud. La idea principal és definir els grups de forma que es minimitzi la variació dins de cada grup al mateix temps que s'intenta maximitzar la variació entre grups diferents [10, 54]. Es poden trobar dues situacions diferents a l'hora de realitzar aquesta classificació depenent de si es fa servir coneixement previ (anàlisi d'agrupacions supervisat) o no (anàlisi d'agrupacions no supervisat). La diferència entre les dues estratègies es mostra a la Figura 2.21.

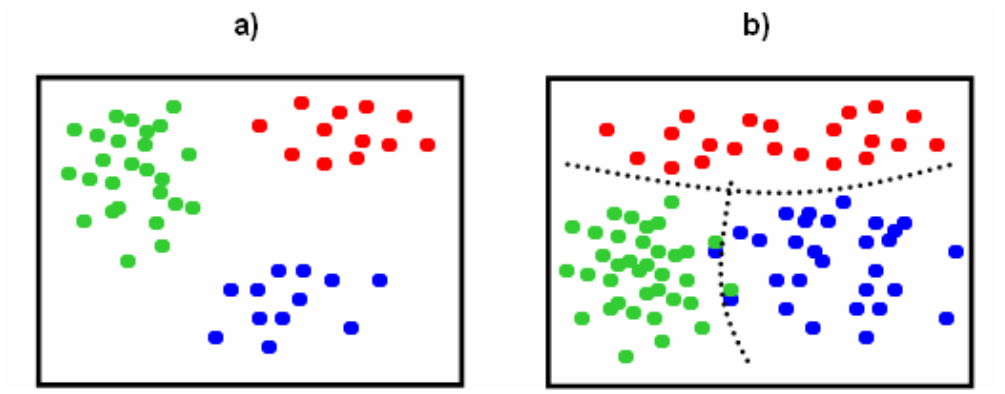


Figura 2.21. Representació esquemàtica dels diferents tipus d'anàlisi d'agrupacions. a) Anàlisi agrupacions no supervisat i b) Anàlisi d'agrupacions supervisat.

En el primer cas (esquerra de la figura), es mostra un exemple d'anàlisi d'agrupacions no supervisat. Així, s'intenta agrupar les mostres tenint en compte únicament la seva situació a l'espai, és a dir, les distàncies existents entre els diferents punts. En el segon cas (dreta de la figura), es mostra un exemple d'anàlisi d'agrupacions supervisat. Així, en primer lloc es classifiquen els objectes coneguts i, posteriorment, s'intenta trobar un mètode que serveixi per discriminar entre els diferents tipus d'objectes de la forma més precisa possible. Habitualment s'acostuma a dir que els mètodes supervisats serveixen per classificar les mostres mentre que els mètodes no supervisats es fan servir per agrupar les mostres o els gens.

Anàlisi d'agrupacions supervisat

L'anàlisi d'agrupacions supervisat classifica la major part de les mostres basant-se en informació coneguda prèvia [73]. Així, els mètodes supervisats utilitzen aquesta informació prèvia, que sol ser oferta per un conjunt de dades d'entrenament o calibratge, i una vegada s'aconsegueix la creació d'un bon model, s'utilitza per predir noves mostres. Alguns dels principals mètodes que han estat àmpliament utilitzats a l'àmbit quimiomètric són el mètode de Mínims Quadrats Parcials (*Partial Least Squares, PLS*) [74, 75], l'Anàlisi Discriminant Linear (*Linear Discriminant Analysis, LDA*) [76, 77], les xarxes neuronals artificials (*Artificial Neural Networks, ANN*) [78, 79] o les Màquines Suportades Vectorialment (*Suport Vector Machines, SVM*) [80, 81].

Anàlisi d'agrupacions no supervisat

Mentre l'anàlisi d'agrupacions supervisat necessita informació precisa sobre la classificació d'algunes de les mostres, els mètodes d'anàlisi d'agrupacions no supervisat no disposen d'aquest tipus d'informació. D'aquesta forma, la formació dels grups es farà únicament en base a similituds entre les mostres o els gens de la nostra matriu de dades [73]. Alguns dels mètodes d'anàlisi d'agrupacions no supervisat més utilitzats en l'anàlisi de dades de micromatrius d'ADN són :

<u>Anàlisi d'agrupacions jeràrquic</u>

L'agrupament jeràrquic transforma les distàncies multidimensionals entre objectes d'una matriu de dades en un conjunt de particions jerarquitzades [10, 73, 82]. Aquesta

jerarquització es pot representar gràficament mitjançant un dendrograma en forma d'arbre en el qual cada grup queda inclòs en un grup més gran fins arribar a un únic grup (Figura 2.22.). Entre els algoritmes que realitzen aquesta agrupació jeràrquica es poden diferenciar dues classes:

1) Els algoritmes aglomerants, en els quals s'inicia amb tants grups com objectes i iterativament es redueix el nombre de grups mitjançant la unió dels grups més semblants de mostres fins arribar a un únic grup.

2) Els algoritmes particionals, en el quals s'inicia amb un únic grup i iterativament es divideix aquest grup inicial fins arribar a tenir el màxim nombre possible de grups, és a dir, un nombre de grups igual al nombre d'objectes.

Hi ha diferents tipus de formes de mesurar les distàncies i mètodes d'enllaçar els diferents grups [73]. Per mesurar les distàncies es fa, servir per exemple, la distància euclídea, la distància de Mahalanobis, la distància de Manhattan o el coeficient de correlació de Pearson. Per enllaçar els diferents grups hi ha mètodes com el de l'enllaç simple, l'enllaç complet, l'enllaç mitja, l'enllaç al veí més proper o més llunya o el mètode de Ward.

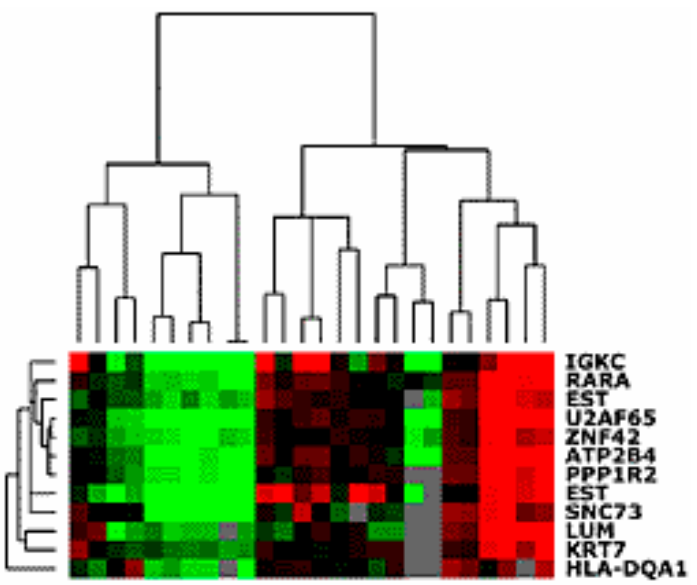


Figura 2.22. Dendrograma aplicat a dades de micromatrius d'ADN adaptada de [70].

Algunes aplicacions d'aquest mètode a dades de micromatrius d'ADN es poden trobar a la literatura [70, 71, 82].

Els mètodes d'anàlisi d'agrupacions jeràrquics van ser en un principi els mètodes més utilitzats per a l'anàlisi de dades de micromatrius d'ADN. L'avantatge més gran és que no necessiten cap tipus d'informació prèvia. Malgrat això, també presenten inconvenients com la dificultat de treballar amb conjunts de dades molt grans a causa dels recursos computacionals necessaris per calcular les matrius de similituds.

Anàlisi d'agrupacions no jeràrquic

Mentre que en el cas de l'anàlisi d'agrupacions jeràrquic es transformen totes les dades experimentals en grups sense necessitat de cap coneixement previ, l'anàlisi d'agrupacions no jeràrquic necessita saber com les dades han de ser particionades [10, 73]. Així, per exemple, es necessita un coneixement previ del nombre de grups present a les dades. Malgrat això, a diferència dels mètodes d'anàlisi supervisats, no es necessita el coneixement previ d'assignació d'una mostra concreta a un grup en particular.

Els principals mètodes d'anàlisi d'agrupacions no jeràrquics es descriuen a continuació.

Mètode *k-means*

El mètode de *k-means* es fa servir àmpliament en l'anàlisi de dades de micromatrius d'ADN ja que proporciona bons resultats basant-se en uns principis molt simples. La idea bàsica és mantenir una estimació del centroide de cadascun dels grups i particionar les dades iterativament de forma que es calculi la suma dels quadrats dels errors en cada punt i, d'aquesta forma, els centroides de cada grup siguin recalculats aconseguint que la suma dels quadrats dels errors sigui mínima [73, 83]. Com a mètode no supervisat, no es necessita conèixer una classificació exacta de les mostres sinó que únicament es necessita definir el nombre de grups i, si és possible, identificar els centroides d'aquests grups.

Així, l'algoritme *k-means* segueix els passos següents:

1) Inicialització amb la definició del nombre de grups (si és possible, es designa un centre per a cadascun d'aquests grups).

2) Assignació de cada objecte al centre del grup més proper, de forma que aquest objecte passa a ser d'aquest grup.
3) Càlcul del nou centre del grup fent la mitjana geomètrica de tots els objectes que pertanyen al grup.
4) Càlcul de la suma dels quadrats dels errors. Si aquest valor no ha millorat en les últimes iteracions s'acaba el procés. Si ha millorat, es torna al segon pas per continuar amb l'optimització iterativa.

El mètode *k-means* s'ha establert com un mètode aconsellable per treballar amb conjunts de dades molt grans degut a la seva complexitat petita que li permet dur a terme un elevat nombre d'iteracions en poc temps. Es poden trobar moltes aplicacions d'aquest mètode a dades de micromatrius d'ADN [84-88]. Malgrat això, presenta inconvenients com ara la incertesa en el càlcul de l'òptim global durant la optimització. Existeix la possibilitat d'utilitzar el mètode *k-means* conjuntament amb tècniques d'algoritmes genètics o de lògica difusa per tal d'assegurar la determinació correcta de l'òptim global.

Anàlisi per components principals (*Principal Component Analysis, PCA*)

L'anàlisi per components principals s'utilitza en l'anàlisi de les dades de micromatrius d'ADN com a eina d'exploració dels possible grups de mostres. Aquest mètode s'explica amb detall al capítol de mètodes quimiomètrics i, a continuació, només es fan unes breus consideracions aplicades a l'anàlisi de dades de micromatrius d'ADN [89-91].

L'anàlisi per components principals descompon la matriu de dades en el producte:

$$\mathbf{D} = \mathbf{U}\,\mathbf{V}^{\mathbf{T}} + \mathbf{E} \qquad \text{Equació 2.8.}$$

En aquesta equació **U** s'anomena matriu de *scores* i proporciona informació sobre la distribució de les mostres en el nou espai vectorial definit pels components principals. $\mathbf{V}^{\mathbf{T}}$ s'anomena matriu de *loadings* i descriu la naturalesa dels components principals (en aquest cas, proporciona informació sobre els perfils d'expressió gènica associats a cada component principal). Les dimensions d'aquestes matrius són **D** (*m*, *n*), **U** (*m*, *nc*)

i $\mathbf{V}^{\mathbf{T}}$ (*nc*, *n*) on *m* és el nombre de mostres, *n* és el nombre de variables (en aquest cas, gens) i *nc* el nombre de components principals seleccionats.

L'anàlisi per components principals intenta capturar la màxima variància de les dades en els components principals que són eixos que expliquen la màxima quantitat de variància i que tenen la propietat de ser ortogonals (no correlacionats) amb altres components principals del mateix conjunt de dades [10, 73]. El nombre de components principals seleccionats pot ser conegut prèviament o pot ser determinat quan el nou component sigui utilitzat per explicar majoritàriament soroll. A partir de la representació de les matrius de *loadings* i de *scores* es podrà obtenir informació sobre els perfils gènics i les agrupacions de les mostres, respectivament. Així, per exemple, en la representació gràfica de la matriu de *scores* es pot determinar el nombre de grups presents en les dades i identificar les mostres que pertanyen a cada grup com es mostra a la Figura 2.23.

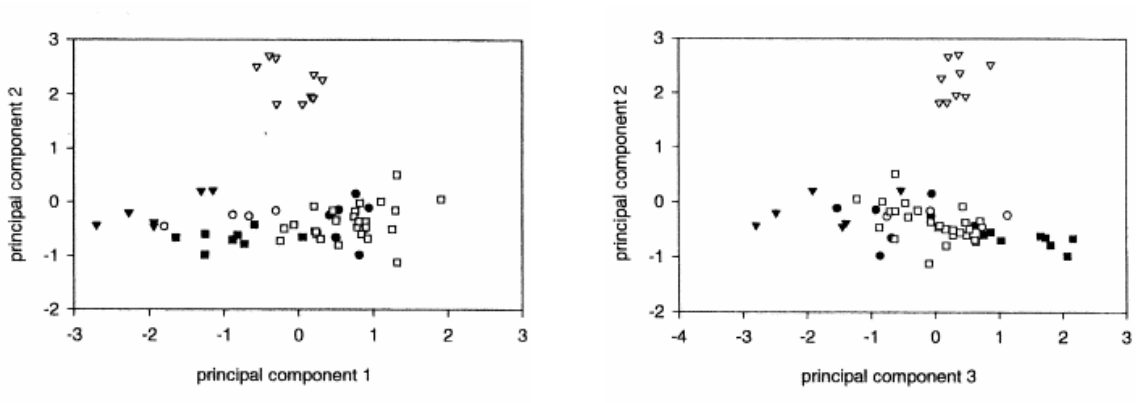


Figura 2.23. Gràfic de *scores* adaptat de [91] on es mostra la separació mitjançant tres components principals de diferents tipus de mostres.

A més del mètode d'anàlisi per components principals, en aquesta Tesi es proposa la utilització del Mètode de Resolució Multivariant de Corbes per Mínims Quadrats Alternats (MCR-ALS) per a l'anàlisi de dades de micromatrius d'ADN. Actualment, únicament es troben referències d'utilització similars per part del grup del Dr. Haaland [92-95]. En aquests treballs, el mètode MCR-ALS s'ha fet servir per a l'anàlisi de dades d'espectroscopia d'imatges hiperespectral que presenten una estructura molt similar a les dades de micromatrius d'ADN. En aquests treballs, MCR-ALS s'ha aplicat per filtrar el senyal de les sondes de DNA del senyal original (que també conté les contribucions del blanc i de les possibles impureses) de forma que es permet la determinació correcta de l'expressió relativa dels gens.

2.5. Bibliografia

(1) Sinden, R. R. (**1994**) *DNA structure and function*. 1st Ed. ed. Academic Press, Londres, UK.

(2) Blackburn, M.; Gait, M. J. (**1990**) *Nucleic Acids in Chemistry and Biology*. ed. Oxford University Press, Oxford, UK.

(3) Watson, J. D.; Crick, F. H. C. (**1953**) Molecular Structure of Nucleic Acids - a Structure for Deoxyribose Nucleic Acid. *Nature*, **171**, 737-738.

(4) Crick, F. (**1970**) Central Dogma of Molecular Biology. *Nature*, **227**, 561-563.

(5) Mathews, C. K.; Holde, K. E. v.; Ahern, K. G. (**2002**) *Bioquímica*. 3era. ed. ed. Pearson Educación: Addison Wesley, Madrid, Espanya.

(6) Johnson, R. S. (**1987**) The Human Genome Project - What Impact on Basic Research. *Faseb Journal*, **1**, 502-505.

(7) Mcgourty, C. (**1989**) Human Genome Project - Dealing with the Data. *Nature*, **342**, 108-108.

(8) Roberts, L. (**1987**) Agencies Vie over Human Genome Project. *Science*, **237**, 486-488.

(9) McKee, T.; McKee, J. R. (**2003**) *Bioquímica. La base molecular de la vida*. 3era. ed. ed. McGraw-Hill / Interamericana, Madrid, Espanya.

(10) Causton, H. C.; Quackenbush, J.; Brazma, A. (**2003**) *Microarray. Gene Expression Data Analysis*. 1era. ed. ed. Blackwell Publishing, Oxford, UK.

(11) Orozco, E.; Gariglio, P. (**1999**) *Genética y biomedicina molecular*. 1era. ed. ed. Uteha: N oriega, Mèxic D.F., Mèxic.

(12) Praseuth, D.; Guieysse, A. L.; Helene, C. (**1999**) Triple helix formation and the antigene strategy for sequence-specific control of gene expression. *Biochimica Et Biophysica Acta-Gene Structure and Expression*, **1489**, 181-206.

(13) Melton, D. A. (**1988**) *Antisense RNA and DNA*. 1era. ed. ed. Cold Spring Harbor Laboratory, Nova York, NY, USA.

(14) Uhlmann, E.; Peyman, A. (**1990**) Antisense Oligonucleotides - a New Therapeutic Principle. *Chemical Reviews*, **90**, 543-584.

(15) *DNA interactive. http://www.dnai.org/timeline/index.html*.

(16) Saenger, W. (**1988**) *Principles of nucleic acid structure*. 2ona. ed. ed. Springer, New York, NY, USA.

(17) Bloomfield, V. A.; Crothers, D. M.; Jr., I. T. (**1999**) *Nucleics Acids. Structure, Properties and Functions*. ed. University Science Books, Sausalito, CA, USA.

(18) Chargaff, E.; Magasanik, B.; Doniger, R.; Vischer, E. (**1949**) The Nucleotide Composition of Ribonucleic Acids. *Journal of the American Chemical Society*, **71**, 1513-1514.

(19) Watson, J. D. (**2004**) *La doble hélice*. 1era. ed. ed. RBA, Barcelona, Espanya.

(20) *Image Library of Biological Macromolecules. http://www.imb-jena.de/IMAGE.html*.

(21) Watson, J. D.; Crick, F. H. C. (**1953**) Genetical Implications of the Structure of Deoxyribonucleic Acid. *Nature*, **171**, 964-967.

(22) Watson, J. D. (**1955**) Biological Consequences of the Complementary Structure of DNA. *Journal of Cellular and Comparative Physiology*, **45**, 109-118.

(23) Crick, F. H. C.; Watson, J. D. (**1954**) The Complementary Structure of Deoxyribonucleic Acid. *Proceedings of the Royal Society of London Series a-Mathematical and Physical Sciences*, **223**, 80-96.

(24) Franklin, R. E.; Gosling, R. G. (**1954**) The Analysis of X-Ray Fibre Diagrams of DNA and the Dependence of Structure on Water Content. *Transactions of the Faraday Society*, **50**, 298-299.

(25) Wilkins, M. H. F.; Hooper, C. W.; Seeds, W. E.; Stokes, A. R.; Wilson, H. R. (**1954**) X-Ray Diffraction Studies of DNA and Nucleoproteins. *Transactions of the Faraday Society*, **50**, 299-299.

(26) *Molecular Structure Laboratory. http://faculty.virginia.edu/molecular-structure/*.

(27) Rinkel, L. J.; Tinoco, I. (**1991**) A Proton Nmr-Study of a DNA Dumbbell Structure with Hairpin Loops of Only 2 Nucleotides-D(Cacgtg-Tgtgcgtgca). *Nucleic Acids Research*, **19**, 3695-3700.

(28) Singh, S.; Patel, P. K.; Hosur, R. V. (**1997**) Structural polymorphism and dynamism in the DNA segment GATCTTCCCCCCGGAA: NMR investigations of hairpin, dumbbell, nicked duplex, parallel strands, and i-motif. *Biochemistry*, **36**, 13214-13222.

(29) Felsenfeld, G.; Davies, D. R.; Rich, A. (**1957**) Formation of a 3-Stranded Polynucleotide Molecule. *Journal of the American Chemical Society*, **79**, 2023-2024.

(30) Felsenfeld, G.; Rich, A. (**1957**) Studies on the Formation of 2-Stranded and 3-Stranded Polyribonucleotides. *Biochimica Et Biophysica Acta*, **26**, 457-468.

(31) Gilbert, D. E.; Feigon, J. (**1999**) Multistranded DNA structures. *Current Opinion in Structural Biology*, **9**, 305-314.

(32) Simonsson, T. (**2001**) G-quadruplex DNA structures - Variations on a theme. *Biological Chemistry*, **382**, 621-628.

(33) *Protein Data Bank http://www.rcsb.org/pdb/*.

(34) Wang, Y.; Patel, D. J. (**1993**) Solution Structure of a Parallel-Stranded G-Quadruplex DNA. *Journal of Molecular Biology*, **234**, 1171-1183.

(35) Escaja, N.; Gelpi, J. L.; Orozco, M.; Rico, M.; Pedroso, E.; Gonzalez, C. (**2003**) Four-stranded DNA structure stabilized by a novel G: C: A: T tetrad. *Journal of the American Chemical Society*, **125**, 5654-5662.

(36) Salisbury, S. A.; Wilson, S. E.; Powell, H. R.; Kennard, O.; Lubini, P.; Sheldrick, G. M.; Escaja, N.; Alazzouzi, E.; Grandas, A.; Pedroso, E. (**1997**) The bi-loop, a new general four-stranded DNA motif. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 5515-5518.

(37) Escaja, N.; Pedroso, E.; Rico, M.; Gonzalez, C. (**2000**) Dimeric solution structure of two cyclic octamers: Four-stranded DNA structures stabilized by A: T: A: T and G: C: G: C tetrads. *Journal of the American Chemical Society*, **122**, 12732-12742.

(38) Mergny, J. L.; Lacroix, L. (**2003**) Analysis of thermal melting curves. *Oligonucleotides*, **13**, 515-537.

(39) *Dna denaturation and annealing. http://courses.cm.utexas.edu/jrobertus/ch339k/overheads-2/ch10_DNAdenat.jpg*.

(40) Michelson, A. M. (**1958**) Hyperchromicity and Nucleic Acids. *Nature*, **182**, 1502-1503.

(41) Fasman, G. D. (**1996**) *Circular Dichroism and the conformational analysis of biomolecules*. 1era. ed. ed. Plenum Press, New York, NY, USA.

(42) Breslauer, K. J. (**1995**) Extracting thermodynamic data from equilibrium melting curves for oligonucleotide order-disorder transitions. *Energetics of Biological Macromolecules*, **259**, 221-242.

(43) Gonzalez, C.; Escaja, N.; Rico, M.; Pedroso, E. (**1998**) NMR structure of two cyclic oligonucleotides. A monomer-dimer equilibrium between dumbbell and quadruplex structures. *Journal of the American Chemical Society*, **120**, 2176-2177.

(44) Xodo, L. E.; Manzini, G.; Quadrifoglio, F.; Vandermarel, G. A.; Vanboom, J. H. (**1991**) Effect of 5-Methylcytosine on the Stability of Triple-Stranded DNA - a Thermodynamic Study. *Nucleic Acids Research*, **19**, 5625-5631.

(45) Xodo, L. E.; Manzini, G.; Alunnifabbroni, M.; Scaggiante, B.; Quadrifoglio, F. (**1992**) Triple-Stranded DNA - Formation, Stability and Application in Biology. *Acta Pharmaceutica*, **42**, 299-307.

(46) Phan, A. T.; Modi, Y. S.; Patel, D. J. (**2004**) Two-repeat Tetrahymena telomeric d(TGGGGTTGGGGT) sequence interconverts between asymmetric dimeric G-quadruplexes in solution. *Journal of Molecular Biology*, **338**, 93-102.

(47) Collins, F. S.; Morgan, M.; Patrinos, A. (**2003**) The human genome project: Lessons from large-scale biology. *Science*, **300**, 286-290.

(48) Arnold, J.; Hilton, N. (**2003**) Genome sequencing - Revelations from a bread mould. *Nature*, **422**, 821-822.

(49) Venter, J. C.; Adams, M. D. *et Al*. (**2001**) The sequence of the human genome. *Journal*, **291**, 1304-1351.

(50) Saiki, R. K.; Scharf, S.; Faloona, F.; Mullis, K. B.; Horn, G. T.; Erlich, H. A.; Arnheim, N. (**1985**) Enzymatic Amplification of Beta-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle-Cell Anemia. *Science*, **230**, 1350-1354.

(51) Saiki, R. K.; Gelfand, D. H.; Stoffel, S.; Scharf, S. J.; Higuchi, R.; Horn, G. T.; Mullis, K. B.; Erlich, H. A. (**1988**) Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA-Polymerase. *Science*, **239**, 487-491.

(52) Chakravarti, A. (**2001**) Single nucleotide polymorphisms. to a future of genetic medicine. *Nature*, **409**, 822-823.

(53) Schena, M. (**2003**) *Microarray Analysis*. 1era. ed. ed. Wiley, Hoboken, NJ, USA.

(54) *DNA Microarrays. http://www.gene-chips.com/*.

(55) van Hal, N. L. W.; Vorst, O.; van Houwelingen, A. M. M. L.; Kok, E. J.; Peijnenburg, A.; Aharoni, A.; van Tunen, A. J.; Keijer, J. (**2000**) The application of DNA microarrays in gene expression analysis. *Journal of Biotechnology*, **78**, 271-280.

(56) Mohr, S.; Leikauf, G. D.; Keith, G.; Rihn, B. H. (**2002**) Microarrays as cancer keys: An array of possibilities. *Journal of Clinical Oncology*, **20**, 3165-3175.

(57) De Bellis, G.; Battaglia, C.; Salani, G.; Bernardi, L. R. (**1999**) Microarray technology in molecular diagnosis and gene expression studies. *Minerva Biotecnologica*, **11**, 227-234.

(58) Ross, J. S.; Schenkein, D. P.; Kashala, O.; Linette, G. P.; Stec, J.; Symmans, W. F.; Pusztai, L.; Hortobagyi, G. N. (**2004**) Pharmacogenomics. *Advances in Anatomic Pathology*, **11**, 211-220.

(59) Chicurel, M. E.; Dalma-Weiszhausz, D. D. (**2002**) Microarrays in pharmacogenomics - advances and future promise. *Pharmacogenomics*, **3**, 589-601.

(60) Yang, Y.; Blomme, E. A. G.; Waring, J. F. (**2004**) Toxicogenomics in drug discovery: from preclinical studies to clinical trials. *Chemico-Biological Interactions*, **150**, 71-85.

(61) Orphanides, G. (**2003**) Toxicogenomics: challenges and opportunities. *Toxicology Letters*, **140**, 145-148.

(62) Southern, E. M. (**1975**) Detection of Specific Sequences among DNA Fragments Separated by Gel-Electrophoresis. *Journal of Molecular Biology*, **98**, 503-517.

(63) Schena, M.; Shalon, D.; Davis, R. W.; Brown, P. O. (**1995**) Quantitative Monitoring of Gene-Expression Patterns with a Complementary-DNA Microarray. *Science*, **270**, 467-470.

(64) Sturn, A. (**2000**) *Cluster Analysis for Large Scale Gene Expression Studies*. ed. Master Thesis at Graz University of Technology.

(65) Duggan, D. J.; Bittner, M.; Chen, Y. D.; Meltzer, P.; Trent, J. M. (**1999**) Expression profiling using cDNA microarrays. *Nature Genetics*, **21**, 10-14.

(66) Brazma, A.; Robinson, A.; Cameron, G.; Ashburner, M. (**2000**) One-stop shop for microarray data - Is a universal, public DNA-microarray database a realistic goal? *Nature*, **403**, 699-700.

(67) Ball, C. A.; Sherlock, G.; Parkinson, H.; Rocca-Serra, P.; Brooksbank, C.; Causton, H. C.; Cavalieri, D.; Gaasterland, T.; Hingamp, P.; Holstege, F.; Ringwald, M.; Spellman, P.; Stoeckert, C. J.; Stewart, J. E.; Taylor, R.; Brazma, A.; Quackenbuch, J. (**2002**) Standards for Microarray data. *Science*, **298**, 539-539.

(68) Brazma, A.; Hingamp, P.; Quackenbush, J.; Sherlock, G.; Spellman, P.; Stoeckert, C.; Aach, J.; Ansorge, W.; Ball, C. A.; Causton, H. C.; Gaasterland, T.; Glenisson, P.; Holstege, F. C. P.; Kim, I. F.; Markowitz, V.; Matese, J. C.; Parkinson, H.; Robinson, A.; Sarkans, U.; Schulze-Kremer, S.; Stewart, J.; Taylor, R.; Vilo, J.; Vingron, M. (**2001**) Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nature Genetics*, **29**, 365-371.

(69) Wentzell, P. D.; Karakach, T. K. (**2005**) DNA microarrays: is there a role for analytical chemistry? *Analyst*, **130**, 1331-1336.

(70) Ross, D. T.; Scherf, U.; Eisen, M. B.; Perou, C. M.; Rees, C.; Spellman, P.; Iyer, V.; Jeffrey, S. S.; Van de Rijn, M.; Waltham, M.; Pergamenschikov, A.; Lee, J. C. E.; Lashkari, D.; Shalon, D.; Myers, T. G.; Weinstein, J. N.; Botstein, D.; Brown, P. O. (**2000**) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, **24**, 227-235.

(71) Golub, T. R.; Slonim, D. K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J. P.; Coller, H.; Loh, M. L.; Downing, J. R.; Caligiuri, M. A.; Bloomfield, C. D.; Lander, E. S. (**1999**) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

(72) Soares, M. B. (**1997**) Identification and cloning of differentially expressed genes. *Current Opinion in Biotechnology*, **8**, 542-546.

(73) Massart, D. L.; Buydens, L. M. C.; Vandegiste, B. G. M. (**1997**) *Handbook of Chemometrics and Qualimetrics*. 1st edn. ed. Elsevier, Amsterdam, The Netherlands.

(74) Nguyen, D. V.; Rocke, D. M. (**2002**) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39-50.

(75) Wold, S.; Martens, H.; Wold, H. (**1983**) The Multivariate Calibration-Problem in Chemistry Solved by the Pls Method. *Lecture Notes in Mathematics*, **973**, 286-293.

(76) Cho, J. H.; Lee, D.; Park, J. H.; Kim, K.; Lee, I. B. (**2002**) Optimal approach for classification of acute leukemia subtypes based on gene expression data. *Biotechnology Progress*, **18**, 847-854.

(77) Fisher, R. (**1936**) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **VII**.

(78) Selaru, F. M.; Xu, Y.; Yin, J.; Zou, T.; Liu, T. C.; Mori, Y.; Abraham, J. M.; Sato, F.; Wang, S.; Twigg, C.; Olaru, A.; Shustova, V.; Leytin, A.; Hytiroglou, P.; Shibata, D.; Harpaz, N.; Meltzer, S. J. (**2002**) Artificial neural networks distinguish among subtypes of neoplastic colorectal lesions. *Gastroenterology*, **122**, 606-613.

(79) Hoskins, J. C.; Himmelblau, D. M. (**1988**) Artificial Neural Network Models of Knowledge Representation in Chemical-Engineering. *Computers & Chemical Engineering*, **12**, 881-890.

(80) Simek, K.; Fujarewicz, K.; Swierniak, A.; Kimmel, M.; Jarzab, B.; Wiench, M.; Rzeszowska, J. (**2004**) Using SVD and SVM methods for selection, classification, clustering and modeling of DNA microarray data. *Engineering Applications of Artificial Intelligence*, **17**, 417-427.

(81) Pontil, M.; Verri, A. (**1998**) Properties of support vector machines. *Neural Computation*, **10**, 955-974.

(82) Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D. (**1998**) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 14863-14868.

(83) Forgy, E. W. (**1965**) Cluster Analysis of Multivariate Data - Efficiency Vs Interpretability of Classifications. *Biometrics*, **21**, 768.

(84) Park, T.; Yi, S. G.; Lee, S.; Lee, S. Y.; Yoo, D. H.; Ahn, J. I.; Lee, Y. S. (**2003**) Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics*, **19**, 694-703.

(85) Dougherty, E. R.; Barrera, J.; Brun, M.; Kim, S.; Cesar, R. M.; Chen, Y. D.; Bittner, M.; Trent, J. M. (**2002**) Inference from clustering with application to gene-expression microarrays. *Journal of Computational Biology*, **9**, 105-126.

(86) Shannon, W.; Culverhouse, R.; Duncan, J. (**2003**) Analyzing microarray data using cluster analysis. *Pharmacogenomics*, **4**, 41-52.

(87) Shapshak, P.; Duncan, R.; Torres-Munoz, J. E.; Duran, E. M.; Minagar, A.; Petito, C. K. (**2004**) Analytic approaches to differential gene expression in aids versus control brains. *Frontiers in Bioscience*, **9**, 2935-2946.

(88) Lee, S. I.; Batzoglou, S. (**2003**) Application of independent component analysis to microarrays. *Genome Biology*, **4**, R66.

(89) Wang, Z. Y.; Wang, Y.; Lu, J. P.; Kung, S. Y.; Zhang, J. Y.; Lee, R.; Xuan, J. H.; Khan, J. V. (**2003**) Discriminatory mining of gene expression microarray data. *Journal of Vlsi Signal Processing Systems for Signal Image and Video Technology*, **35**, 255-272.

(90) Verhoeckx, K. C. M.; Bijlsma, S.; de Groene, E. M.; Witkamp, R. F.; van der Greef, J.; Rodenburg, R. J. T. (**2004**) A combination of proteomics, principal component analysis and transcriptomics is a powerful tool for the identification of biomarkers for macrophage maturation in the U937 cell line. *Proteomics*, **4**, 1014-1028.

(91) Crescenzi, M.; Giuliani, A. (**2001**) The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data. *Febs Letters*, **507**, 114-118.

(92) Haaland, D. M.; Timlin, J. A.; Keenan, M. R.; Jones, H. D. T.; Stork, C. L.; Melgaard, D. K.; Sinclair, M. B. (**2004**) Multivariate analysis of hyperspectral images for biotechnology applications. *Abstracts of Papers of the American Chemical Society*, **228**, U113-U113.

(93) Martinez, M. J.; Aragon, A. D.; Rodriguez, A. L.; Weber, J. M.; Timlin, J. A.; Sinclair, M. B.; Haaland, D. M.; Werner-Washburne, M. (**2003**) Identification and removal of contaminating fluorescence from commercial and in-house printed DNA microarrays. *Nucleic Acids Research*, **31**, e18.

(94) Timlin, J. A.; Haaland, D. M.; Sinclair, M. B.; Aragon, A. D.; Martinez, M. J.; Werner-Washburne, M. (**2005**) Hyperspectral microarray scanning: impact on the accuracy and reliability of gene expression data. *BMC Genomics*, **6**, 72.

(95) Haaland, D. M.; Timlin, J. A.; Sinclair, M. B.; Benthem, M. H. v.; http://www.genomes2life.org/publications/SPIE-4959-06-final.pdf, **2004**.