Universitat Ramon Llull

TESI DOCTORAL

Títol	**Desenvolupament del programari ArIS (*Artificial Intelligence Suite*): implementació d'eines de cribratge virtual per a la química mèdica**
Realitzada per	Roger Tejedor Estrada
en el Centre	Escola Tècnica Superior IQS
i en el Departament	Química Orgànica
Dirigida per	Dr. Jordi Teixidó i Closa i Dr. Santi Nonell Marrugat

C.Claravall, 1-3
08022, Barcelona
Tel. 936 022 200
Fax. 936 022 249
E-mail: urlsc@sec.url.es
www.url.es

Nihil Magnum Nisi Bonum
No hi ha grandesa sense bondat

Agraïments

Aquest treball s'ha realitzat en el laboratori de Disseny Molecular del Grup d'Enginyeria Molecular (Institut Químic de Sarrià, Universitat Ramon Llull), sota la direcció del Dr. Jordi Teixidó i Closa i el Dr. Santi Nonell Marrugat, als quals vull agrair l'oportunitat de poder-lo dur a terme. Gràcies també a l'IQS i a la Generalitat de Catalunya per les beques predoctorals de les quals he gaudit en els darrers anys.

Durant la tesi he tingut l'ocasió de col·laborar amb un gran nombre de companys i amics. Molta de la feina que es presenta a continuació no s'hauria pogut realitzar sense la seva valuosa ajuda: J. Iglesias en la implementació de les xarxes neuronals predictives, J. Bernard en la implementació dels mètodes de *pruning* i el càlcul d'absorcions bifotòniques, N. Sabaté en la mesura experimental del logP de fotosensibilitzadors, X. Tomàs en l'anàlisi estadística de les regressions de logP, A. Serrano en la mesura experimental de formulacions químiques, V. Sauri i L. Serrano-Andrés (e.p.d.) en l'estudi teòric de sistemes porficènics, M. Duran en les mesures experimentals i en el recolzament computacional en l'estudi del 9ATPPo, M. Riera en la implementació dels SOM, ... sense oblidar el suport tècnic i ajuda existencial del departament de TICS, especialment de D. Serrano, S. Arasa i J. Vericat.

De forma independent, han sorgit altres col·laboracions amb companys de dins i fora del grup de recerca als quals, tot i no figurar en la memòria, vull agrair la confiança dipositada: A. Monteagudo i M. Pérez en l'estudi de propietats de tensioactius, E. Cirera en l'establiment de mètodes QSRR, E. Roig i C. Ribas en la predicció de paràmetres cromatogràfics, J. Vidal en la predicció de pKa, O. Planas en l'estudi fotofísic de nous porficens, l'IUCT en la selecció de quimioteques combinatòries, R. Moltó en la simulació de membranes biològiques, J. Menacho i L. González en l'estudi de metodologies per a la innovació docent, i (gairebé) I. Pérez en el disseny de caspases.

En el camp més personal, voldria agrair l'oportunitat que molts m'han brindat per descobrir la docència: el deganat de l'escola tècnica superior IQS, J. Cuadros per deixar-me col·laborar en els laboratoris d'informàtica i de càlcul numèric, el departament de comunicació per creure que un químic teòric també pot entrar al laboratori, i molt especialment i sincera a J. Menacho i J. Teixidó per haver-me donat l'oportunitat de poder gaudir de les classes de química computacional i matemàtiques durant els darrers 6 anys. És clar que, sense tots els alumnes, res hauria estat possible ...

El reguitzell de models, línies de codi i fitxers generats durant aquest període no tindrien gens de sentit (i m'atreveixo a dir, de valor) a no ser per totes aquelles persones que m'han acompanyat i recolzat en tot moment. Agraeixo l'amistat, consell i companyonia d'autèntics mestres, com són (en ordre alfabètic, per no fer distincions) A. Balfagón, L. Comellas, J. Menacho, J.J. Molins i J. Teixidó; les rialles i cabòries dels companys del grup de recerca i amics del laboratori de Disseny Molecular, amb en Miquel i en Rubén, i la paciència de l'Àlex i en Marc en el tram final de la redacció. Gràcies als de dins i fora del grup de recerca i als de dins i fora del químic, amb una menció especial a en Xavier, l'Eulàlia, en Sebi, la Laura, la Laiona i en Toni, la Núria i en Sergi, la Marta, en Quim i l'Iris, ... tots pessigollers pompeuencs menjadors de canelons de Sant Esteve.

Finalment, voldria mostrar l'agraïment més sincer a la meva família, especialment a la Ma. Teresa, a la Núria i a la Laia. El més sincer i absolut possible, senzillament per tot: podria fer una llista *incommensurable* amb tot el que els vull agrair, però encara em quedaria curt.

Roger

Sumari

El disseny molecular de sistemes d'interès per a la química mèdica i per al disseny de fàrmacs sempre s'ha trobat molt lligat a la disponibilitat sintètica dels resultats. Des del moment que la química combinatòria s'incorpora dins de l'esquema sintètic, canvia el paper que ha de jugar la química computacional: la diversitat d'estructures possibles a sintetitzar fa necessària la introducció de mètodes, com el cribratge virtual, que permetin avaluar la viabilitat de grans quimioteques virtuals amb un temps raonable.

Els mètodes quimioinformàtics responen a la necessitat anterior, posant a l'abast de l'usuari mètodes eficaços per a la predicció teòrica d'activitats biològiques o propietats d'interès. Dins d'aquests destaquen els mètodes basats en la relació quantitativa d'estructura-activitat (QSAR). Aquests han demostrat ser eficaços per l'establiment de models de predicció en l'àmbit farmacològic i biomèdic. S'ha avaluat la utilització de mètodes QSAR no lineals en la teràpia fotodinàmica del càncer, donat que és una de les línies de recerca d'interès del Grup d'Enginyeria Molecular (GEM) de l'IQS. El disseny de fotosensibilitzadors es pot realitzar a partir de la predicció de propietats fisicoquímiques (com l'espectre d'absorció i la hidrofobicitat del sistema molecular), i de l'estudi de la seva localització subcel·lular preferent, la qual ha demostrat recentment jugar un paper molt important en l'eficàcia del procés global.

Per altra banda, les xarxes neuronals artificials són actualment un dels mètodes més ben valorats per a l'establiment de models QSAR no lineals. Donat l'interès de disposar d'un programari capaç d'aplicar aquests mètodes i que, a més, sigui prou versàtil i adaptable com per poder-se aplicar a diferents problemes, s'ha desenvolupat el programari ArIS. Aquest inclou els principals mètodes de xarxes neuronals artificials, per realitzar tasques de classificació i predicció quantitativa, necessaris per a l'estudi de problemes d'interès, com és la predicció de l'activitat anti-VIH d'anàlegs de l'AZT, l'optimització de formulacions químiques o el reconeixement estructural de grans sistemes moleculars.

Sumario

El diseño molecular de sistemas de interés para la química médica y para el diseño de fármacos siempre ha estado condicionado por la disponibilidad sintética de los resultados. Desde el momento en que la química combinatoria se incorpora en el esquema sintético, cambia el papel de la química computacional: la diversidad de estructuras que pueden sintetizarse hace necesaria la introducción de métodos, como el cribado virtual, que permitan evaluar la viabilidad de grandes quimiotecas virtuales en un tiempo razonable.

Los métodos quimioinformáticos responden a la necesidad anterior, ofreciendo al usuario métodos eficaces para la predicción teórica de actividades biológicas o propiedades de interés. Entre ellos destacan los métodos basados en la relación cuantitativa de estructura-actividad (QSAR), que han demostrado ser eficaces para establecer modelos de predicción en el ámbito farmacológico y biomédico. Se ha evaluado la utilización de métodos QSAR no lineales en terapia fotodinámica del cáncer, dado que es una de las líneas de investigación de interés del *Grup d'Enginyeria Molecular* (GEM) del IQS. El diseño de fotosensibilizadores se puede realizar a partir de la predicción de propiedades fisicoquímicas (como su espectro de absorción o su hidrofobicidad) y del estudio de su localización subcelular preferente, la cual ha demostrado recientemente jugar un papel muy importante en la eficacia del proceso global.

Por otro lado, las redes neuronales artificiales son actualmente uno de los métodos mejor valorados para establecer modelos QSAR no lineales. Es por ello que resulta muy interesante disponer de un programa capaz de aplicar estos métodos y que, además, sea lo suficientemente versátil y adaptable como para poder aplicarse a distintos problemas, según las necesidades del usuario. Por este motivo se ha desarrollado el programa ArIS, el cual incluye los principales métodos de redes neuronales artificiales para realizar tareas de clasificación y predicción cuantitativa, necesarios para el estudio de problemas de interés como la predicción de la actividad anti-VIH de análogos del AZT, la optimización de formulaciones químicas o el reconocimiento estructural de grandes sistemas moleculares.

Summary

Molecular modelling of interesting systems for medicinal chemistry and drug design highly depends on availability of synthetic results. Since combinatorial chemistry was incorporated into the synthetic scheme, the role of computational chemistry has changed: the structural diversity of candidates to be synthesized requires the introduction of computational methods which are able to screen large virtual libraries.

Answering to this requirement, chemoinformatics offers many kinds of different methods for predicting biological activities and molecular properties. One of the most relevant techniques among them is Quantitative Structure-Activity Relationships (QSAR), which can be used to establish prediction models for both, pharmacological and biomedical sectors. The use of non-linear QSAR methods has been evaluated in photodynamic therapy of cancer, one of the research areas of the *Grup d'Enginyeria Molecular* (GEM) at IQS. Molecular design of photosensitizers can be performed by computational studies of their physicochemical properties (absorption spectra or hydrophobicity, for example) and subcellular localization, which becomes a key factor in the efficacy of the overall process.

Furthermore, artificial neural networks are nowadays rated as one of the very best methods for establishing non-linear QSAR models. Developing software that includes all these methods would be certainly interesting. Implemented algorithms should be versatile and easily adaptable for their use in any problems. We have developed ArIS software, which includes the most important methods of artificial neural networks for classification and quantitative prediction. ArIS has been used to predict anti-HIV activity of AZT-analogues, for optimization of chemical formulations and for structural recognition in large molecular systems, among others.

Índex

Abreviatures

a.u.	*Atomic units,* unitats atòmiques
ADALINE	*adaptive linear neuron*
AG	aparell de Golgi
ANN	*artificial neural networks,* xarxes neuronals artificials
BBB	*blood brain barrier,* barrera hematoencefàlica
BH3	*Bcl2 homology-3 domain*
BP	*back-propagation learning rule*
DA	*discriminant analysis,* anàlisi discriminant
DFT	*density functional theory,* teoria del funcional de la densitat
DIF	localització difosa
DISC	*death-inducing signaling complex*
DNA	àcid desoxirribonucleic
eq	equació
FS	fotosensibilitzador
FS-BE	*forward selection-backward elimination*
GA	*genetic algorithm,* algorisme genètic
GEM	grup d'enginyeria molecular
GM	unitats Göpert-Mayer
GNN	*genetic neural network,* xarxa neuronal genètica
HF	*Hartree-Fock*
HOMO	*highest occupied molecular orbital,* últim orbital molecular ocupat

HPLC	*high pressure liquid chromatography,* cromatografia líquida d'alta pressió
IA	intel·ligència artificial
k-NN	*k-nearest neighbors*
LBDD	*ligand-based drug design,* disseny de fàrmacs basat en els lligands
LDA	*linear discriminant analysis,* anàlisi discriminant lineal
LDL	*low density lipoprotein,* lipoproteïna de baixa densitat
LIS	lisosoma
LMS	*least-mean-square*
LOO	*leave-one-out*
LUMO	*lowest unoccupied molecular orbital,* primer orbital molecular desocupat
MAb	*monoclonal antibody,* anticòs monoclonal
MADALINE	*multiple adaptive linear neuron*
MC	membranes intracel·lulars
MEM	membrana
MIT	mitocondri
MM	*molecular mechanics,* mecànica molecular
MPTP	*mitochondrial permeability transition pore*
MRII	*MADALINE rule II*
NNTree	*neural network tree,* arbre de xarxes neuronals
NUC	nucli cel·lular
OTH	altres localitzacions subcel·lulars
PBR	*peripheral benzodiacepine receptor,* receptor perifèric de benzodiacepina
PC	*principal component,* component principal
PCA	*principal components analysis,* anàlisi de components principals
PCM	*polarized continuum method*
PDT	*photodynamic therapy,* teràpia fotodinàmica
PLR	*perceptron learning rule*

PLS	*partial least squares*, mínims quadrats parcials
PRALINS	*Program for Rational Analysis in Silico*
PTPC	*permeability transition pore complex*
QM	*quantum mechanics*, mecànica quàntica
QSAR	*quantitative structure-activity relationship*, relació quantitativa estructura-activitat
QSPR	*quantitative structure-property relationship*, relació quantitativa estructura-propietat
QSRR	*quantitative structure-retention relationship*, relació quantitativa estructura-retenció
RE	reticle endoplasmàtic
RMSE	*root mean square error*, error quadràtic mitjà
ROS	*reactive oxygen species*, espècies reactives de l'oxigen
SAR	*structure-activity relationships*, relacions estructura-activitat
SBDD	*structure-based drug design*, disseny de fàrmacs basat en l'estructura
SCF	*self-consistent field*, camp autocoherent
SOM	*self-organizing maps*
TD-DFT	*time-dependent DFT*, DFT depenent del temps
TSA	*tumor-specific antigen*, antigen específic de cèl·lules tumorals
UV	ultraviolat
VAPS	valors propis
VEPS	vectors propis
ZPE	*zero point energy*, energia en el punt zero

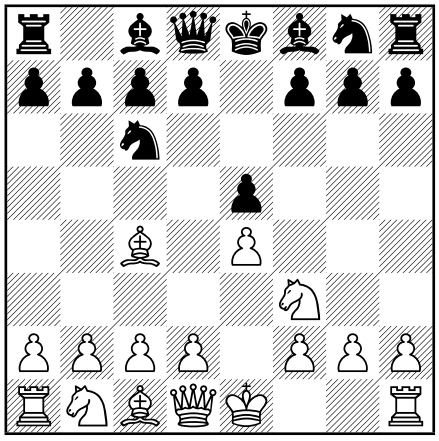
Índex de figures

Índex de taules

Índex d'algorismes

Part I

Introducció



El *giuoco piano* és una de les apertures obertes més utilitzades en escacs i la més antiga. Permet un joc tranquil i metòdic amb el qual desenvolupar el joc de totes les peces de forma esglaonada. Amb el temps se n'han creat nombroses variacions (algunes d'elles força atrevides) com el gambet Evans, en el qual les blanques ofereixen el sacrifici d'un peó per tal d'afavorir el seu desenvolupament. En aquests casos, així com en l'etapa final de la partida (on després de l'enfrontament directe entre els dos exèrcits sovint queden només unes poques peces), cal tenir un coneixement molt gran de les possibilitats que ofereix cadascuna de les peces de què es disposa. Per sort, cada jugador les coneix abans d'iniciar el joc.

Aquesta sort no la tingueren els primers químics de la història, que es van haver d'enfrontar amb el joc de l'experimentació química sense el coneixement previ de les peces: l'estructura de la matèria.

La contemplació de l'entorn que ens envolta ha despertat des de sempre la curiositat humana, suscitant la voluntat de saber què el forma. Les primeres teories sobre la constitució de la matèria s'establiren a l'antiga Grècia. Entre els segles VII i IV aC, el gran interès científic que mostraren els hel·lènics féu passar de la concepció que tot està format únicament per quatre elements (terra, foc, aire i aigua) a la idea primigènia de l'àtom de Demòcrit.

S'estaven començant a descobrir les peces del joc, i la partida no es va fer esperar. El xoc cultural entre Grècia i Egipte va confluir en l'art de la *khemeia* (l'alquímia), on el saber sobre l'alteració de les substàncies es matisava amb creences religioses. Durant aquest període es varen crear el que es podria considerar els primers laboratoris, tot i el misticisme i la mala reputació dels alquimistes (que portà a l'encriptació del saber, on les substàncies s'associaren a cossos celestes). Perseguint la transmutació de la matèria, l'experimentació a cegues va fer que s'iniciés l'ús de tècniques com la destil·lació o l'escalfament al bany maria, que actualment són considerades tècniques bàsiques.

La química representà fins ben entrat el segle XVII la voluntat de comprendre la naturalesa humana a través de l'experimentació directa. En l'època en què els científics no es trobaven limitats per l'estigma de la seva professió, les aportacions interdisciplinàries de científics com Boyle, amb l'estudi dels gasos ideals, encetaren el camí de química física. Reapareixia de nou l'estudi de la matèria però des d'un punt de vista diferent, on els mètodes químics i les teories físiques es recolzaven per avançar conjuntament a través del complex mètode científic.

A principis del segle XX la recerca en física fou aclaparadora, i els seus fonaments trontollaren sota la teoria general de la relativitat d'Einstein i la teoria quàntica d'Schrödinger. Aquesta última permetia estudiar el comportament de la matèria des de la perspectiva dels seus components més elementals (com els electrons). Donat que moltes de les reaccions químiques es poden descriure com la formació i trencament d'enllaços, la teoria quàntica introduí el marc teòric necessari per a la descripció teòrica de la matèria i la seva utilització en l'àmbit químic. A més, la simplicitat de la seva formulació fa que el seu resultat hagi estat fins i tot inclòs dins dels llibres de text.

La culminació de la supèrbia matemàtica en una equació:

$$i\hbar\frac{\partial \Psi_t}{\partial t} = \widehat{H}\Psi_t$$

Capítol 1

La física del món molecular

Seguint la rutina de cada diumenge, el tren de les 16:35h surt amb cinc minuts de retard de l'estació de Manlleu. S'agraeix poder aixoplugar-se de la pluja que, tossudament, continua caient. Al seu pas per la Garriga, de camí a Barcelona, el tren avança a velocitat constant v i en línia recta. En un moment determinat, cauen dos llamps (A i B) que són simultanis respecte a la via ferroviària. Ho seran també respecte el tren?

Com cada any des del 2002, Terrassa celebra a la primavera la Fira Modernista. Entre els artesans que s'hi apleguen, hom hi pot trobar el senyor Antoni, de Breda, que comparteix el seu art fet de ceràmica amb la multitud de nens que l'envolten tots els dies que dura la mostra, mostrant-se sempre simpàtic i atent. Els nens que no estan fent cua per una de les seves petites obres d'art, estan bocabadats davant del taller de joguines d'un artesà estranger, el qual els mostra un petit artefacte (que anomena giroscopi) que s'aguanta misteriosament sobre una corda, sense caure. Com pot ser que aquest objecte pugui precessionar sobre la punta del seu dit índex sense que li caigui?

La física aporta resposta a les preguntes plantejades anteriorment: la teoria de la relativitat especial d'Einstein[1] permet assegurar que la resposta a la primera pregunta ha de ser necessàriament negativa, mentre que el concepte de moment angular permet descriure el comportament del giroscopi. I és que la física ens permet descriure fidelment la realitat macroscòpica que coneixem, mesurar-la i fins i tot predir-la. Però en el món microscòpic les regles són diferents i les teories que podem aplicar en els esdeveniments del nostre dia a dia no poden ser emprats a nivell molecular. Per això va caldre desenvolupar una nova física: la **mecànica quàntica**.

La mecànica quàntica (*Quantum Mechanics*, QM) és la part de la física que aporta el marc teòric necessari per a l'estudi de l'estructura més íntima de la natura, permetent ésser tractada matemàticament i així poder predir les seves propietats.

En ser pronunciat, el propi mot provoca sovint una certa hostilitat (l'autor encoratja fermament a fer la prova), possiblement deguda a l'estesa creença sobre la seva hermeticitat matemàtica. Per aquest motiu, en aquest treball es pren la llibertat de simplificar la càrrega teòrica de les següents

seccions, on s'intenta explicar de forma planera els mètodes que s'empren en la discussió del treball experimental.

1.1 L'essència matemàtica de la matèria

1.1.1 Un món de boles i molles

La representació gràfica de molècules mitjançant un determinat nombre de línies (segons l'ordre d'enllaç) que uneixen un conjunt de nodes, pot conduir, tot i haver estudiat l'enllaç químic, a què hom s'imagini els compostos químics com un conjunt d'àtoms (reduïts a boles) que es troben units físicament entre sí mitjançant una vareta. Donada la dificultat de crear-se una representació mental de les molècules, lluny d'aquesta concepció errònia i mancada de realisme, tot comportament s'intenta descriure a partir d'ella. Així doncs, sabent que els àtoms no són estàtics, sinó que cada sistema molecular té un conjunt de modes normals de vibració, la visió més simplista de l'enllaç químic correspon a un sistema de boles i molles, figura 1.1.

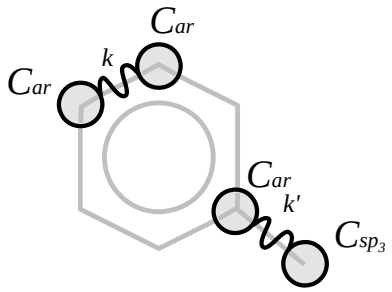


Figura 1.1: *Representació esquemàtica de la molècula de toluè per mecànica molecular. S'indica el tractament de dos dels seus enllaços, segons una constant de força i indicant el tipus d'àtoms.*

La mecànica molecular recull aquesta idea, que aplicada de forma rigorosa i controlada, pot utilitzar-se en benefici de la descripció puntual o al llarg del temps de grans sistemes moleculars com proteïnes o àcids nucleics.

Sota aquesta perspectiva, la definició d'un sistema molecular es realitza per mitjà d'un camp de forces (*forcefield*), que inclou la descripció geomètrica del sistema (mitjançant $3N$-6 coordenades internes corresponents a les $3N$ coordenades cartesianes menys 6 que són redundants per translació i rotació) i els paràmetres energètics que contribueixen a la definició de l'energia interna de la molècula.

Donat que el sistema es troba regit per la constant de força dels enllaços i per la naturalesa de cada àtom (*atom type*), l'energia es defineix com una combinació dels diferents moviments que poden patir els àtoms i dels efectes no enllaçants.

$$E = E_{\text{str}} + E_{\text{ben}} + E_{\text{tor}} + E_{\text{oop}} + E_{\text{non}} \tag{1.1}$$

E_{str} Fa referència a l'energia corresponent a la tensió d'enllaç (*stretching*). Normalment s'utilitza la corba de Morse per definir-la, ja que ha de presentar un comportament proper a l'harmònic entorn a la posició d'equilibri i evidenciar la dissociació atòmica a valors allunyats.

E_{ben} Terme corresponent a la flexió o modificació de la posició d'equilibri d'un angle format entre dos enllaços (*bending*). Se simula com una variació de la llei de Hooke.

E_{tor} Situació anàloga a l'anterior però avaluant la modificació d'angles díedres (*torsion*). Els perfils d'energia per a la rotació d'enllaços ha de presentar periodicitat i normalment es defineixen mitjançant series de Fourier.

E_{oop} El terme *out of plane* penalitza la modificació de la planarietat de determinades zones (com per exemple en dobles enllaços C=C).

E_{non} La contribució de les forces intermoleculars (*non-bonded interactions*) recullen bàsicament l'efecte de les interaccions electrostàtiques i de van der Waals.

Tots aquests paràmetres (i alguns més, corresponents a altres termes energètics com la formació de ponts d'hidrogen) es troben definits dins del *forcefield* per cadascun dels diferents tipus d'àtom possibles. Per aquest motiu és molt important assegurar la correcta descripció del sistema i l'assignació que es realitza per a cada *atom type*.

Cada *forcefield* conté una parametrització diferent dels termes anteriors, que en estar basats en valors experimentals dependran en gran mesura del sistema a tractar. Això significa que abans d'utilitzar un determinat *forcefield* cal assegurar que aquest permetrà definir correctament les característiques del sistema en estudi. Serveixi d'exemple la comparació entre els camps de forces AMBER[2] i PEF95SAC:[3] AMBER és un *forcefield* especialment parametritzat per a proteïnes i àcids nucleics, però no permet descriure correctament molècules petites. En canvi, el PEF95SAC està especialment dissenyat per ser utilitzat en la descripció d'hidrocarburs que continguin en la seva estructura únicament àtoms de carboni, hidrogen i oxigen.

Tot i procurar transmetre una visió realista dels sistemes macromoleculars, la mecànica molecular és incapaç d'oferir un tractament dels processos que impliquen un reordenament electrònic, com ara la formació i el trencament d'enllaços. En aquests casos, cal recórrer a mètodes que permetin descriure de forma acurada el comportament més íntim de la matèria: la mecànica quàntica.

1.1.2 Els postulats de la mecànica quàntica

1. L'estat d'un sistema en un instant determinat queda totalment definit per una funció d'ona (Ψ).

2. Tot observable té associat un operador (hermític, que garanteix l'obtenció d'un valor propi real) que actua sobre Ψ.

3. La funció d'ona Ψ que defineix l'estat del sistema no es troba determinada unívocament. La probabilitat d'obtenir el valor a_i en mesurar l'observable A quantitzat, correspon a una expressió que depèn de les seves funcions pròpies (Φ_{ij}):

$$P_\Psi(A \rightarrow a_i) = \sum_{j=1}^{d_i} |\langle \Phi_{ij} | \Psi \rangle|^2 \tag{1.2}$$

4. La mesura d'un observable en un instant t comporta un canvi en l'estat actual del sistema, sempre i quan Ψ no sigui funció pròpia de l'observable. La mesura obté com a resultat un valor propi de l'operador (a_i), i deixa el sistema en l'estat definit pel vector propi corresponent:

$$\Psi = \phi_i \tag{1.3}$$

5. L'evolució temporal de Ψ està determinada per l'equació d'Schrödinger depenent del temps:

$$i\hbar \frac{\partial \Psi}{\partial t} = \widehat{H}\Psi \tag{1.4}$$

Interpretació dels postulats

En comparació amb la física clàssica, l'acceptació de la funció d'ona com a descripció total de l'estat d'un sistema, és un canvi de paradigma fonamental (taula 1.1). La dualitat ona–partícula proposada per Einstein com a explicació de l'efecte fotoelèctric (posada de manifest per l'experiment de la doble escletxa), portà a De Broglie (1923) a suggerir la seva generalització considerant també la dualitat partícula–ona:[4] si les ones (llum) poden tractar-se com partícules (fotons), es podrien descriure les partícules com a ones?

Taula 1.1: *Algunes diferències entre la física associada a sistemes clàssics i quàntics.*

Sistemes clàssics	**Sistemes quàntics**
Sempre és possible precisar millor les condicions inicials de cada experiment	Conèixer Ψ en el moment de realitzar la mesura significa disposar de la màxima informació possible sobre el seu estat (postulat 1)
Es pot calcular la probabilitat d'obtenir un resultat a partir de la repetició consecutiva d'un experiment	La primera mesura de totes deixa el sistema en un estat propi de l'observable mesurat, de manera que (per moltes vegades que es repeteixi de forma consecutiva) sempre s'obtindrà aquell mateix valor (postulat 4)
Les lleis de Newton determinen l'evolució d'un sistema a partir d'unes condicions inicials: $x(0) \longrightarrow x(t)$	L'evolució temporal de Ψ queda completament determinada un cop es coneix la seva expressió en un instant determinat (postulat 5): $\Psi_0 \longrightarrow \Psi_t$
Qualsevol propietat física d'un sistema es troba determinada independentment del seu estat dinàmic	Un observable d'un sistema es troba determinat si i només si l'estat del sistema és propi de l'operador corresponent

La hipòtesi de De Broglie fou demostrada experimentalment anys més tard per Davisson i Germer (1927), que demostraren que un feix d'electrons es difracta com si fos un paquet d'ones. Per tant, en l'estudi de partícules elementals, hom pot apel·lar a la seva naturalesa ondulatòria per a descriure-les.

En aquest sentit, es proposa la funció d'ona (Ψ) definida en el postulat 1, la qual ha de contenir la informació de la posició de les n partícules del sistema al llarg del temps:

$$\Psi = \Psi\left(x, y, z, t\right) \tag{1.5}$$

Es defineix l'espai vectorial $\{\Psi\}$ de les funcions d'ona que pot prendre un sistema, sobre el qual es defineix el producte intern segons l'equació 1.6, i en conjunt té estructura de l'espai de Hilbert ($\mathscr{H}$).

$$\langle\Psi|\Psi\rangle = \int_{\mathbb{R}^{3n}} \Psi^* \Psi \, d\, r \tag{1.6}$$

Encara que Ψ no té significat físic, sí que el té el seu quadrat: Bohr interpreta Ψ^2 com la funció de densitat de la probabilitat. Això implica que, com que les partícules han d'estar en algun lloc, Ψ hagi d'estar normalitzada:

$$\langle\Psi|\Psi\rangle = 1 \tag{1.7}$$

Disposant de l'expressió pel producte intern (eq. 1.6), queda definida la manera com mesurar el sistema. Segons el postulat 2, a tot observable li correspon una aplicació lineal ($\widehat{A}$), anomenat operador quàntic, que és hermític i té l'espai de Hilbert com a espai d'origen i final:

$$\widehat{A}: \mathscr{H} \rightarrow \mathscr{H} \tag{1.8}$$

Segons el principi de correspondència, a partir de l'expressió clàssica d'un observable (a) es pot trobar l'operador quàntic anàleg ($\widehat{A}$) i les seves funcions (estats) pròpies, ϕ:

$$\widehat{A}\Phi = a_i\Phi \tag{1.9}$$

El conjunt de les funcions pròpies d'un operador hermític constitueixen una base ortonormal, $\{\Phi\}$. Això significa que qualsevol funció de l'espai Ψ (com per exemple la funció pròpia d'un operador qualsevol) es pot expressar com a combinació lineal d'aquestes:

$$\Psi = \sum_{k=1}^{N} c_k\Phi_k \tag{1.10}$$

Així doncs, segons el postulat 3, en mesurar l'observable A hom pot obtenir qualsevol dels valors propis (VAPS) admesos. Quan l'espectre d'A és discret, és a dir, quan només hi ha uns valors d'a possibles (determinats pels nombres quàntics), la probabilitat d'obtenir el valor a_i (el qual pot presentar un ordre de degeneració D_i) correspon a l'eq. 1.2. Atenent al resultat de l'eq. 1.10, l'expressió per a la probabilitat es pot reescriure com:

$$P_\Psi(A \rightarrow a_i) = \sum_{j=1}^{D_i} |c_{i,j}|^2 \tag{1.11}$$

Hom pot adonar-se que la mesura de l'observable, segons l'eq. 1.11, suposa la projecció de Ψ sobre el conjunt $\{\Phi\}$ amb valor propi a_i (i.e. el subespai propi d'$\widehat{A}$ corresponent al valor a_i). Aquesta projecció comporta un resultat inconcebible des del punt de vista clàssic: la mesura d'un observable té com a conseqüència l'alteració impredictible de l'estat del sistema (postulat 4). Després de la mesura, el valor de l'observable queda totalment determinat, fent que l'estat passi a ser un estat propi d'$\widehat{A}$, que pot ser qualsevol d'ells amb la probabilitat definida en l'eq. 1.11.

D'altra banda, el valor esperat de l'operador ($\langle A\rangle$) es pot calcular com el valor mitjà dels observables:

$$\langle A\rangle = \left\langle \Psi|\widehat{A}|\Psi \right\rangle = \frac{\int \Psi^*\widehat{A}\Psi\, d\,r}{\int \Psi^*\Psi\, d\,r} \tag{1.12}$$

Per tal d'avaluar com varia en el temps la funció d'ona d'un sistema entre dues mesures consecutives cal recórrer a l'equació d'Schrödinger dependent del temps, corresponent al postulat 5 (eq. 1.4, pàg. 38).

L'operador implicat en aquesta expressió és l'operador Hamiltonià, $\widehat{H}$ (que rep el nom en honor a les equacions del moviment de Hamilton, s.XIX, per a la descripció de sistemes formats per un gran nombre de partícules[5,6]). $\widehat{H}$ és l'operador quàntic que descriu l'energia clàssica del sistema, i ha de contenir tots els termes necessaris. A més de l'energia cinètica i potencial (que per a l'objecte d'estudi d'aquest treball és suficient considerar funció només de les coordenades espacials) se li poden afegir

altres termes referents a processos que poden afectar al sistema com l'acoblament spin-òrbita o la interacció amb una radiació electromagnètica.

Inicialment es considera el cas d'un sistema no relativista conservatiu ($\widehat{H}$ no depèn explícitament del temps), mentre que el tractament de sistemes que evolucionen en el temps, sotmesos a radiació electromagnètica, es comenten en la secció 1.3, pàg. 46. En aquest escenari, l'equació d'Schrödinger es redueix a una equació diferencial de variables separables, eq. 1.13.

$$i\hbar\frac{\partial\Psi(r,t)}{\partial t} = i\hbar\psi(r)\frac{df(t)}{dt} = e^{-i\frac{E\cdot t}{\hbar}}\widehat{H}\psi(r) \tag{1.13}$$

Si se separa la part independent del temps, l'expressió que s'obté correspon a la que es coneix comunament com equació d'Schrödinger (independent del temps).

$$\widehat{H}\Psi = E\Psi \tag{1.14}$$

1.1.3 Construint el castell de cartes

La definició de les energies cinètica i potencial de l'Hamiltonià es pot descompondre en components nuclears i electrònics. D'aquesta manera, mentre l'energia cinètica (E_c) es defineix amb només dos components (E_c^n per a la nuclear i E_c^e per a l'electrònica), l'energia potencial ha de considerar l'atracció entre nuclis i electrons (E_p^{n-e}), la repulsió entre nuclis (E_p^{n-n}) i la repulsió electrònica(E_p^{e-e}, anomenada també energia de correlació):

$$\widehat{H} = E_c^n + E_c^e + E_p^{n-e} + E_p^{n-n} + E_p^{e-e} + \ldots \tag{1.15}$$

L'aproximació de Born i Oppenheimer

Donat que la quantitat de moviment dels nuclis és molt superior a la dels electrons, es pot considerar que aquests últims es mouen suficientment de pressa com perquè la funció d'ona dels electrons s'adapti instantàniament a un petit moviment dels nuclis. Això significa que es pot desacoblar el moviment dels electrons al dels nuclis, fent que es pugui definir un Hamiltonià electrònic ($\widehat{H}_e$, per cada una de les posicions del nuclis) i un de nuclear ($\widehat{H}_n$).

L'equació d'Schrödinger electrònica (eq. 1.16) permet descriure la superfície d'energia potencial del sistema, ja que la funció potencial efectiu (E') depèn de les coordenades nuclears, però no de la seva velocitat.

$$\widehat{H}_e\Psi_e = E'\Psi_e \tag{1.16}$$

Per altra banda, l'equació nuclear (on $\widehat{H}_n = E_c^n + E'$) permet descriure el moviment dels nuclis, així com els seus estats vibracionals.

Teorema variacional

L'equació 1.16 només es pot solucionar de forma analítica per a sistemes monoelectrònics ($E_p^{e-e}=0$). Per a la resta de sistemes, cal trobar la millor Ψ que descriu el sistema. Per aquest motiu, el teorema variacional estableix que el valor d'energia esperada, $\langle E'\rangle$, calculada amb una funció d'ona de prova (Φ), ha de ser necessàriament superior a l'energia de l'estat fonamental (E^0). Així doncs, la millor Ψ que podem trobar serà aquella que minimitzi el valor $\langle E'\rangle$.

$$\langle E'\rangle = \frac{\left\langle \Phi|\widehat{H}_e|\Phi\right\rangle}{\langle\Phi|\Phi\rangle} \geqslant E^0 \tag{1.17}$$

Existeixen diferents mètodes per proposar de forma sistemàtica diferents definicions de Φ. Un exemple d'ells és el mètode Rayleigh-Ritz, on es proposa un conjunt finit de funcions de base, $\{\psi_k\}$, i on els coeficients de la combinació lineal $\{c_k\}$(eq. 1.18) es calculen com els valors propis (VAPS) de l'equació secular matricial obtinguda d'aplicar la condició de mínim, eq. 1.19 ($H_e\mathbb{C}=E'\mathbb{SC}$).

$$\Phi = \sum_k c_k\psi_k \tag{1.18}$$

$$\frac{\partial E'}{\partial c_k}=0 \tag{1.19}$$

El camp autocoherent

Apel·lant al teorema variacional, la resolució de l'equació d'Schrödinger electrònica proposada pel mètode del camp autocoherent (*Self-Consistent Field*, SCF), o mètode de Hartree-Fock (HF), es basa en un procés iteratiu on es les funcions d'ona aproximades que es proven corresponen a funcions monoelectròniques (orbitals).

Si l'equació d'Schrödinger electrònica fos de variables separables, és a dir, si s'obviés el terme de repulsió electrònica (E_p^{e-e}), $\widehat{H}_e$ es podria definir com la suma d'un conjunt d'Hamiltonians monoelectrònics (eq. 1.20). Aleshores, Ψ_e es podria escriure com un producte de funcions d'ona monoelectròniques normalitzades (φ_i), i rep el nom de funció d'ona de Hartree (eq. 1.21).

$$\widehat{H}_e = \sum_i \widehat{h}_i \tag{1.20}$$

$$\Psi_e = \prod_i \varphi_i(r_i) \tag{1.21}$$

El fet que tots els electrons siguin iguals i tots estiguin a tot arreu, fa que la funció d'ona hagi de ser forçosament antisimètrica per tal d'assegurar que no hi són alhora. Dit d'una altra manera, la funció d'ona de Hartree (eq. 1.20) no compleix el principi de Pauli i cal assegurar que dos electrons (fermions) no tinguin els mateixos nombres quàntics fent que la funció d'ona sigui antisimètrica respecte l'intercanvi de coordenades. Per aquest motiu es defineix la funció d'ona com un determinant,

que rep el nom de determinant d'Slater:

$$\Psi_e = \frac{1}{\sqrt{N!}} \begin{vmatrix} \varphi_1(r_1) & \varphi_2(r_1) & \dots & \varphi_n(r_1) \\ \varphi_1(r_2) & \varphi_2(r_2) & \dots & \varphi_n(r_2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1(r_n) & \varphi_2(r_n) & \dots & \varphi_n(r_n) \end{vmatrix} \tag{1.22}$$

Les funcions monoelectròniques $\{\phi_i\}$ reben el nom d'orbitals moleculars i es poden descriure com combinació lineal de funcions matemàtiques de base, $\{\chi_k\}$, per facilitar el càlcul:

$$\varphi_i = \sum_k c_{ki}\chi_k \tag{1.23}$$

L'aproximació de Hartree-Fock. Sigui com sigui, la repulsió electrònica s'ha de tenir en compte en la definició de l'Hamiltonià. Per tal que la inclusió d'aquest efecte no entri en conflicte amb la descripció monoelectrònica, es realitza l'aproximació de Hartree-Fock. Aquesta considera que cada electró es mou en un camp mitjà generat per la resta d'electrons i els nuclis. El moviment dels electrons entre ells es considera independent i, en considerar-los indistingibles entre sí, tots els orbitals es troben ocupats per tots els electrons (bo i que no simultàniament, eq. 1.22). Així, el càlcul de les integrals bielectròniques es realitza de forma aproximada tenint en compte l'electró en qüestió i una mitjana de la resta.

Com es pot tenir en compte, però, l'efecte mitjà dels electrons si no s'han calculat prèviament? Davant d'aquest problema, el mètode SCF planteja un procés iteratiu. El procés es podria iniciar amb un conjunt de funcions d'ona de Hartree, corresponents a un conjunt de solucions aproximades, que s'utilitzen per estimar la correlació electrònica (integrals d'intercanvi i de Coulomb). El valor obtingut s'empra seguidament per trobar un nou conjunt. Així, SCF refina de manera gradual les solucions monoelectròniques, que es corresponen en energies cada vegada més baixes (segons el mètode variacional), fins a assolir l'estat en què els resultats obtinguts per a cadascun dels electrons es mantenen invariables.

Donat que els orbitals calculats analíticament per l'hidrogen són els únics que han demostrat tenir existència real, la primera aproximació fou intentar utilitzar orbitals hidrogenoides tipus Slater (STO) com a funcions de base. Així doncs, en el mètode SCF es considera que els orbitals moleculars es representen com combinació lineal d'orbitals atòmics (CLOA-OM).[7] Malauradament aquestes funcions tenen una derivada discontinua i són difícilment integrables, motiu pel qual s'utilitzen combinacions de funcions Gaussianes que simulin la funció de base STO. Per exemple, la base 6-31G* utilitzada en la part III (pàg. 145), utilitza 6 funcions Gaussianes per definir les funcions STO dels electrons de les capes internes, 3 Gaussianes per a la primera sèrie d'electrons de valència i 1 Gaussiana per a la segona sèrie de valència. A més, l'asterisc indica que s'afegeixen funcions de polarització superiors a tots els àtoms pesants (definint orbitals p).

Limitacions del mètode HF. En ser un mètode basat en el teorema variacional i en contemplar la correlació electrònica només de manera amitjanada, el mètode HF acabarà assolint un valor asimp-

tòtic (anomenat límit de HF, E_{HF}) superior a l'energia real no relativista (E^0), encara que el valor de l'energia disminueixi en cada iteració. La diferència energètica correspondrà precisament a l'energia de correlació E_{corr}, figura 1.2.

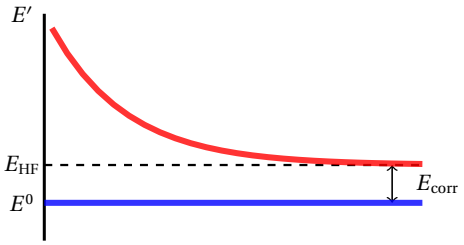


Figura 1.2: *Representació esquemàtica del límit energètic del mètode HF.*

1.1.4 Mètodes post-HF

Els mètodes post Hartree-Fock sorgeixen de la voluntat de tractar la repulsió instantània entre tots els electrons, en comptes de descriure-la de forma amitjanada (com fan els mètodes HF), i poder superar així el límit d'energia.

Teoria de pertorbacions

Aquesta metodologia permet trobar solucions aproximades d'$\widehat{H}$ a partir de la descripció d'un sistema semblant i de solució coneguda ($\widehat{H}_0$):

$$\widehat{H} = \widehat{H}_0 + \widehat{H}' \tag{1.24}$$

La pertorbació ($\widehat{H}'$) pot jugar diferents rols segons es defineixi: pot correspondre a la part responsable de no tenir solució analítica, a la part de l'Hamiltonià dependent del temps o a l'efecte d'un camp extern sobre els estats estacionaris coneguts d'un sistema.

Les funcions pròpies (o estats) originals no pertorbades constitueixen una base, de manera que la solució de la pertorbació es pot trobar com una combinació lineal dels estats no pertorbats. Aquesta solució s'expressa com un desenvolupament en sèrie de potències (correccions de diferent ordre) que modifica el valor inicial:

$$\Psi_n = \Psi_n^0 + \lambda^1 \Psi_n^I + \lambda^2 \Psi_n^{II} \ldots \tag{1.25}$$

$$E_n = E_n^0 + \lambda^1 E_n^I + \lambda^2 E_n^{II} \ldots \tag{1.26}$$

Per tal de superar el límit de HF, el mètode Moller-Plesset[8] (MPn, sent n l'ordre de la correcció i equivalent a HF per a $n = 1$) inclou dins de la part pertorbacional la correlació electrònica no amitjanada, considera el tractament dels orbitals virtuals i renuncia a ser variacional. De forma orientativa, es considera que el mètode MP2 (corresponent a la correcció de segon ordre i emprat en la secció 7.3, pàg. 176) permet corregir el 50% de l'energia de correlació electrònica i el seu ús s'ha estès fins i tot a macromolècules.[9]

Teoria del funcional de la densitat

Els mètodes basats en la teoria del funcional de la densitat (*Density Functional Theory*, DFT) consideren que les propietats dels sistemes en estat fonamental es poden calcular a partir de la densitat electrònica[10] (teorema de Hohenberg-Kohn[11,12]). Això significa que, sota aquest punt de vista, un observable (que en mecànica quàntica és $\langle A\rangle = \langle\Psi|A|\Psi\rangle$) passa a ser calculat com una funció del quadrat de la funció d'ona: $\langle A\rangle = f\left(\Psi^2\right)$.

Aquesta nova interpretació de Ψ^2 com $\langle\Psi|\Psi\rangle$, té una important repercussió sobre el càlcul de l'energia, fent que cadascun dels termes que la composen s'hagi de referenciar a la densitat electrònica, i calgui transformar l'expressió de l'energia en un funcional.

El funcional a utilitzar en DFT no ha de permetre només obtenir el valor de l'energia en ser aplicat sobre Ψ^2, sinó també contemplar l'energia de correlació. Malauradament no existeix cap base teòrica que permeti deduir la seva expressió. Per aquest motiu es recorre a funcionals aproximats, que poden estar parametritzats amb valors definits empíricament.

Durant la realització d'aquest treball, els càlculs que impliquen mètodes DFT s'han portat a terme mitjançat el funcional híbrid B3LYP.[13,14] Els funcionals híbrids permeten complementar (com a combinació lineal) els resultats dels càlculs HF amb els obtinguts per un o més mètodes DFT referents a les energies de correlació i d'intercanvi. En el cas concret del funcional B3LYP, el component de l'energia d'intercanvi i correlació es calcula a partir de l'energia HF i les aproximacions obtingudes per quatre funcionals diferents (LDA, B88, VWN3 i LYP). La resolució iterativa d'aquest tipus d'equacions es pot realitzar, de forma anàloga al mètode SCF, mitjançant un sistema de Kohn-Sham.

1.2 Mètodes semiempírics

El cost computacional dels mètodes basats en la mecànica quàntica es pot rebaixar (sense arribar a la simplificació extrema de la mecànica molecular) amb mètodes semiempírics. Aquests mètodes consideren tan sols els electrons de València, apel·lant a la baixa influència dels electrons de les capes internes sobre el comportament químic del sistema. Això es tradueix en la reducció de l'efecte dels electrons interns a un mer apantallament del camp elèctric nuclear que arriba als electrons de València.

Els mètodes semiempírics, a més, redueixen el nombre d'integrals a resoldre negligint les energies associades a la repulsió coulòmbica i a l'intercanvi electrònic. Aquest fet implica la incorporació de paràmetres experimentals que pal·liïn aquesta mancança i garanteixin l'obtenció de resultats coherents. El tipus de parametritzacions i la seva naturalesa determinen els diferents tipus de mètodes semiempírics. El mètode ZINDO (emprat a la part III), correspon a un mètode semiempíric derivat del mètode INDO,[15] especialment parametritzat pel tractament d'estats excitats.[16] Per la definició dels estats fonamentals, actualment s'utilitzen els mètodes derivats de PM3[17] i AM1.[18]

1.3 Interacció radiació–matèria

Dins dels mètodes de càlcul quàntics, s'aprofundeix intencionadament en la manera com es pot simular la interacció entre la radiació (r) i la matèria (m), donat que el seu tractament és clau en la predicció d'espectres d'absorció (capítol 7, pàg. 147).

Des del punt de vista físic, un sistema format per càrregues en moviment genera un camp elèctric ($\overline{F}$) capaç d'interaccionar amb el camp d'una radiació electromagnètica ($\overline{B}$). Aquesta interacció (E_{r-m}), pot ser descrita en funció dels moments dipolars (elèctric, $\overline{d}$, i magnètic, $\overline{\mu}$) i multipolars (Q_{ii}), eq.1.27.

$$E_{r-m} = V_0 q - \overline{F_0}\cdot\overline{d} - \frac{1}{2}\left[\left(\frac{\partial F_x}{\partial x}\right)_0 Q_{xx} + \left(\frac{\partial F_y}{\partial y}\right)_0 Q_{yy} + \ldots + \left(\frac{\partial F_z}{\partial z}\right)_0 Q_{zz}\right] - \overline{B_0}\cdot\overline{\mu} - \ldots \tag{1.27}$$

La descripció quàntica dels sistemes sotmesos a l'efecte de la radiació requereix la resolució de l'equació d'Schrödinger depenent del temps, donat que els camps elèctric i magnètic varien segons aquesta variable. Tal com s'ha començat a apuntar en la secció anterior, la radiació electromagnètica es pot definir com una pertorbació dins de l'expressió de l'Hamiltonià,[19] figura 1.3.

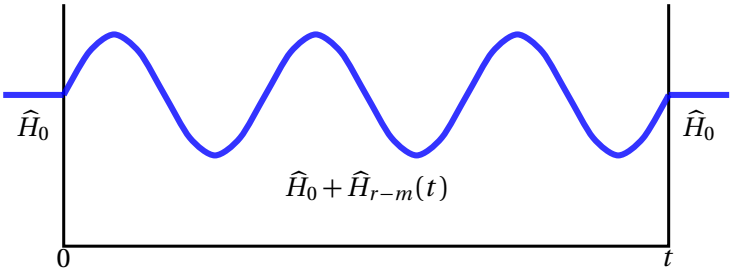


Figura 1.3: *Representació esquemàtica de la teoria de pertorbacions. En presència de radiació electromagnètica, el sistema es troba pertorbat amb un $\widehat{H}_{r-m}(t)$ que depèn del temps.*

Així doncs, l'expressió de l'Hamiltonià canviarà durant el temps de radiació:

$$\widehat{H}_{r-m} = V_0\widehat{q} - \overline{F_0}\cdot\widehat{\overline{d}} - \frac{1}{2}\left[\left(\frac{\partial F_x}{\partial x}\right)_0 \widehat{Q_{xx}} + \left(\frac{\partial F_y}{\partial y}\right)_0 \widehat{Q_{yy}} + \ldots + \left(\frac{\partial F_z}{\partial z}\right)_0 \widehat{Q_{zz}}\right] - \overline{B_0}\cdot\widehat{\overline{\mu}} - \ldots \tag{1.28}$$

Per estrany que sembli, cadascun dels termes de l'eq. 1.28 pot ser calculat matemàticament. Per una banda, els camps elèctric i magnètic es poden calcular segons les equacions 1.29 (corresponents a una ona plana, monocromàtica i polaritzada) i per l'altra, les expressions dels moments multipolars es poden calcular substituint els operadors quàntics dins les expressions clàssiques.

$$\widehat{F} = \widehat{u_x}\cdot F\cdot\cos\left(\omega t - kz + \varphi\right) \tag{1.29}$$

$$\widehat{B} = \widehat{u_y}\cdot B\cdot\cos\left(\omega t - kz + \varphi\right) \tag{1.30}$$

Les solucions de l'estat pertorbat es poden calcular de forma aproximada a partir de combinaci-

ons lineals dels estats estacionaris coneguts independents del temps. En conèixer les funcions pròpies de $\widehat{H}_0$, $\{\Phi_k\}$, i en ser aquestes base del mateix espai de Hilbert al qual pertany $\Psi(x,t)$, hom pot calcular la funció d'ona del sistema pertorbat com una combinació lineal on el component temporal recau exclusivament sobre els coeficients c_i:

$$\Psi(x,t) = \sum_k a_{k_i}(t)\Phi_k(x) = \sum_k \left(c_{k_i}(t) \cdot e^{-i\frac{E_k t}{\hbar}} \right) \Phi_k(x) \tag{1.31}$$

La determinació de la funció $\Psi(x,t)$ es redueix al càlcul dels coeficients $c_{k_i}(t)$ (el subíndex i fa referència a l'estat inicial $\Phi_i(x) \cong \Psi(x,0)$). Per aquest motiu s'introdueix l'expressió de la funció d'ona dependent del temps dins l'equació d'Schrödinger depenent del temps (eq. 1.4):

$$i\hbar\frac{\partial}{\partial t}\left(\sum_k c_{k_i}(t) \cdot e^{-i\frac{E_k t}{\hbar}} \Phi_k(x) \right) = \left(\widehat{H}_0 + \widehat{H}_{r-m}(t) \right) \cdot \left(\sum_k c_{k_i}(t) \cdot e^{-i\frac{E_k t}{\hbar}} \Phi_k(x) \right) \tag{1.32}$$

Aplicant l'operador derivada a la part esquerra de la igualtat seguint la regla de la cadena, i atenent al fet que l'equació d'Schrödinger independent del temps es continua complint, hom pot obtenir el següent resultat:

$$i\hbar \sum_k \frac{\partial c_{k_i}}{\partial t} e^{-i\frac{E_k t}{\hbar}} \Phi_k(x) = \sum_k c_{k_i}(t) \cdot e^{-i\frac{E_k t}{\hbar}} \widehat{H}_{r-m}(t)\Phi_k(x) \tag{1.33}$$

Donat que el conjunt de funcions pròpies d'$\widehat{H}_0$ es poden escollir ortonormals, hom pot multiplicar ambdós costats de la igualtat per una funció $\Phi_f(x)$ tal que $\left\langle \Phi_f | \Phi_k \right\rangle = \delta_{fk}$, obtenint l'expressió corresponent a la variació dels coeficients al llarg del temps:

$$\frac{\partial c_{f_i}}{\partial t} = \frac{1}{i\hbar} \sum_k c_{k_i}(t) e^{i(E_f - E_k)t} \left\langle \Phi_f | \widehat{H}_{r-m}(t) | \Phi_k \right\rangle \tag{1.34}$$

El conjunt de totes les equacions diferencials, que es poden obtenir per a cadascuna de les funcions Φ_f que pertanyen a $\widehat{H}_0$, formen un sistema d'equacions.

El mètode de les pertorbacions dependents del temps

El sistema d'equacions diferencials obtingut segons l'eq. 1.34 està acoblat, de manera que no té una solució exacta. Per aquest motiu es considera que els efectes de la interacció són poc importants i que els coeficients c_{k_i} varien poc durant aquest temps.

Desacoblant el sistema s'obté una equació diferencial resoluble en variables separables per cada coeficient. Integrant i aplicant com a condició inicial $c_{k_i}(t) \sim c_{k_i}(0) = \delta_{ki}$, hom obté un valor aproximat de cada coeficient.

El mètode de les pertorbacions dependents del temps proposa la millora successiva del valor aproximat de c_{k_i} substituint-lo dins de l'equació 1.34. En resoldre aquesta nova expressió, s'obté una aproximació de segon ordre per als valors dels coeficients. Aquest procés es pot repetir de manera successiva per millorar aquests valors.

Predient l'espectre d'absorció

Una vegada s'han calculat els coeficients c_{k_i}, es disposa de la definició de la funció d'ona que regeix el sistema durant la pertorbació electromagnètica, eq. 1.31. El fet de descriure l'estat pertorbat com una combinació lineal d'estats estacionaris implica que en la seva definició hi poden participar estats amb diferent energia. Per tant, tot i que el sistema torna a ser conservatiu després de la pertorbació, durant el procés deixa de ser estacionari: la funció d'ona deixa de ser pròpia de l'Hamiltonià. Això porta a parlar de la probabilitat d'assolir un nou estat. Segons el tercer postulat hom pot calcular la probabilitat d'obtenir un determinat valor d'energia E_f:

$$P_t(E_f) = |c_{f_i}(t)|^2 \tag{1.35}$$

Es diu que s'ha produït una transició radiativa entre els estats Φ_i i Φ_j quan $E_i \neq E_f$ i $P_t(E_f) \neq 0$. En aquest cas, la probabilitat $P_t(E_f)$ correspon a la probabilitat de la transició $\Phi_i \rightarrow \Phi_j$.

Així doncs, la forma de l'espectre d'absorció d'un determinat sistema molecular es pot deduir a partir de les probabilitats de les transicions calculades.

La teoria del funcional de la densitat dependent del temps

De forma anàloga, es pot aplicat la pertorbació $\widehat{H}_{r-m}$ dins del formulisme DFT. En aquests casos es parla de *Time-Dependent* DFT o TD-DFT.

1.4 Tocant de peus a terra

Encara que la teoria quàntica permet descriure de forma acurada els sistemes químics i moleculars, la seva aplicació sempre s'ha trobat lligada a les eines de càlcul disponibles.

La introducció dels ordinadors dins de l'escena de la química teòrica va permetre revolucionar l'ús de la quàntica per a la descripció molecular de sistemes petits, i no tan petits a mesura que la informàtica ha anat evolucionant. Tot i així, la modelització acurada de sistemes amb més d'un centenar d'àtoms amb mecànica quàntica continua sent un repte que es tradueix en elevats temps de càlcul.

La informàtica, en aquest sentit, juga un paper doble: per un costat dóna el recolzament necessari per poder resoldre les complexes expressions matemàtiques pròpies de la quàntica, i alhora obre el camí a altres tècniques que puguin aprofitar l'alta velocitat de processament. Aquesta situació condueix a preguntar-se si podrien existir mètodes alternatius que permetin descriure determinades propietats dels sistemes químics, sense haver de recórrer a la mecànica quàntica.

Capítol 2

Què és la quimioinformàtica?

En els darrers anys la química teòrica ha patit una gran compartimentació que ha donat pas a un seguit de disciplines altament especialitzades. Uns dels exemples més clars d'aquesta escissió són la bio i la quimioinformàtica. El poder que exerceixen ambdós prefixos fa que la separació entre aquestes dues disciplines tan intrarelacionades continuï essent debatuda actualment per la comunitat científica.[20] Des de la perspectiva de la química mèdica, l'ampli coneixement de les bases bioquímiques d'una malaltia o disfunció de què es disposa actualment, fa que en ser estudiades, hom estigui gairebé obligat a moure's del món macroscòpic al microscòpic. I en aquest trajecte, *on s'acaba la part biològica i on comença la part química*?

És l'objecte d'estudi el que dictamina a quin món pertany una tècnica computacional. Tal com es pot suposar, aquesta distinció esdevé cada vegada més difosa a mesura que hom s'apropa a la zona d'encontre, figura 2.1[a]. Normalment es considera que aquells mètodes que permeten la simulació de sistemes macromoleculars de naturalesa biològica, típicament proteïnes i àcids nucleics, pertanyen al domini de la bioinformàtica. És relativament comú adaptar expressions com engalzament (en anglès, *docking*), modelització per homologia, dinàmica molecular o alineament de seqüències, dins del llenguatge bioinformàtic. Totes aquestes tècniques permeten estudiar el comportament de macromolècules, és a dir de sistemes formats per milers d'àtoms, i que avui en dia no poden aspirar més que a una descripció global del sistema.

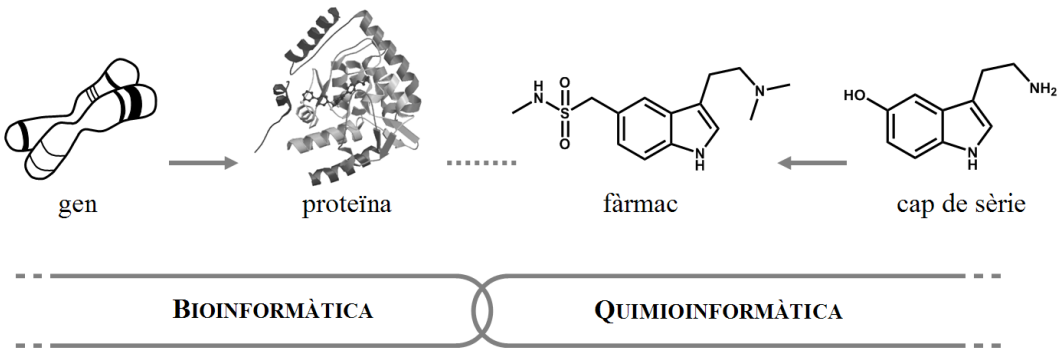


Figura 2.1: *Diagrama il·lustratiu de l'àmbit d'estudi de la bio i quimioinformàtica.*

[a]Adaptació de l'esquema proposat per Engel i Gasteiger

Per la seva banda, la quimioinformàtica se centra més en la naturalesa química del sistema en estudi. Essent conscient que la millor manera de descriure l'essència dels sistemes moleculars és mitjançant la resolució de l'equació d'Schrödinger (capítol 1, pàg. 35), i que els mètodes que ho permeten fer solen requerir una important potència i elevats temps de càlcul, la quimioinformàtica es concep com una eina dinàmica i pragmàtica que sovint premia la utilitat dels seus resultats per sobre d'una escrupolosa descripció teòrica. Aquesta diferència respecte a la química quàntica és fonamental per entendre l'extensió del seu ús.

En plantejar alternatives útils i assequibles en un breu període de temps, la química mèdica ha trobat en aquests mètodes l'aliat perfecte per al cribratge virtual de grans quimioteques de forma prèvia a la seva síntesi i fer front, així, a l'esclat combinatori que suposà la incorporació de la química combinatòria dins de l'esquema sintètic. No és d'estranyar, doncs, que la predicció de propietats d'interès (com pot ser l'avaluació de l'activitat d'un candidat a fàrmac front una determinada diana terapèutica) acapari gran part de l'atenció dels estudis quimioinformàtics.

Els mètodes quimioinformàtics que s'apliquen dins l'esquema del disseny de fàrmacs es poden classificar en dos grans grups, segons sigui l'objecte principal d'estudi:

- **Mètodes basats en estructura** (*Structure-based drug design*, SBDD): Per presentar activitat biològica, els lligands han d'interaccionar amb un receptor. Els mètodes basats en estructura se centren en l'estudi del receptor. Es considera, per tant, que el fet de disposar d'un coneixement exhaustiu de la seva estructura és suficient per orientar el disseny de lligands potencialment actius.

  En tractar-se de sistemes macromoleculars (amb un elevat nombre d'àtoms), els mètodes computacionals aplicats han de permetre descriure i operar amb el sistema completa en un temps raonable. Alguns exemples d'aquest tipus de mètodes són:

  - El disseny *De Novo*. Basant-se en la necessitat dels lligands d'establir unes determinades interaccions, el disseny *De Novo* se centra en la creació fragmental de lligands a l'interior del receptor, que es pren com a motlle. Així, es força que els lligands proposats se situïn en una posició favorable per establir les interaccions adients amb el centre actiu. La construcció de les molècules pot seguir diferents esquemes (de dins a fora o de fora cap a dins), però en tot cas cal disposar d'una quimioteca de fragments adequada, que permeti unir les diferents peces amb un criteri químic acceptable i sintèticament accessible.

  - L'engalzament (*Docking*). Estudia de forma dinàmica la interacció entre el lligand i el receptor. Mesurant en cada cas l'energia d'interacció de cada conformació i situació del lligand en el receptor, es pot identificar el centre actiu d'una proteïna i avaluar el seu comportament i efecte en establir-se la interacció. Tot i així, no es disposa d'una única definició per a la funció de puntuació (o d'*scoring*) que s'utilitza com a aproximació de l'energia d'interacció, de manera que els resultats obtinguts poden dependre de la funció utilitzada, figura 2.2.

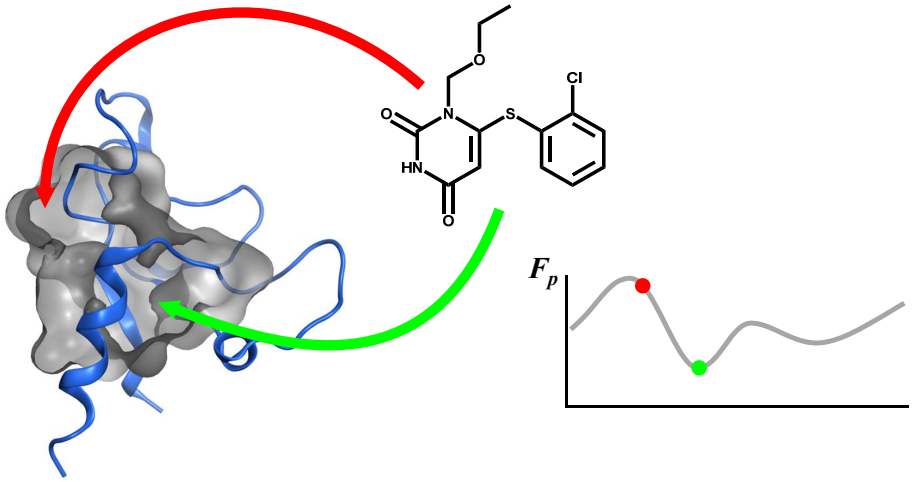


Figura 2.2: *Representació esquemàtica del mètode d'engalzament. Durant el procés de càlcul, el lligand adopta diferents conformacions que permeten que interaccioni amb residus de diferents regions del receptor. La funció de puntuació (F_p) s'avalua en cadascun d'aquests casos, identificant la més estable.*

- **Mètodes basats en els lligands** (*Ligand-based drug design*, LBDD): Utilitzen el coneixement dels lligands per poder-ne proposar d'altres que presentin l'activitat desitjada. Concretament, aquests mètodes permeten identificar les característiques estructurals diferencials d'un conjunt de lligands que presenten activitat front una diana terapèutica. Per aquest motiu cal comparar-los amb un conjunt de lligands no actius, de manera que aquests mètodes utilitzen un gran nombre de compostos, que s'organitzen en forma de quimioteques. Tot i que la descripció quàntica d'aquests sistemes és, per tant, impracticable i cal recórrer a mètodes aproximats, aquests no deixen de tenir en compte les característiques atòmiques.

  Dins d'aquest grup s'hi troben inclosos tots aquells mètodes que utilitzen l'estructura molecular dels lligands per establir models de predicció, com per exemple:

  - Els models farmacofòrics. Són un tipus de mètodes fragmentals que, en comptes d'identificar subestructures (com ho faria un mètode de cerca basat en la semblança), intenten descriure la naturalesa i la situació de les interaccions que fan que un determinat lligand sigui actiu front a un receptor concret. Les característiques considerades estàndards per a la descripció de les interaccions lligand-receptor solen ser la capacitat de formar ponts d'hidrogen (donador o acceptor), el caràcter hidrofòbic i la identificació de zones ionitzades (positivament i negativa). A partir de la inspecció d'un conjunt de lligands actius es pot crear una hipòtesi del mapa farmacofòric que pot ser emprat posteriorment per cribrar molècules candidates, figura 2.3. Alternativament, les característiques farmacofòriques identificades es poden utilitzar a posteriori com a descriptors per establir mètodes QSAR.[21]

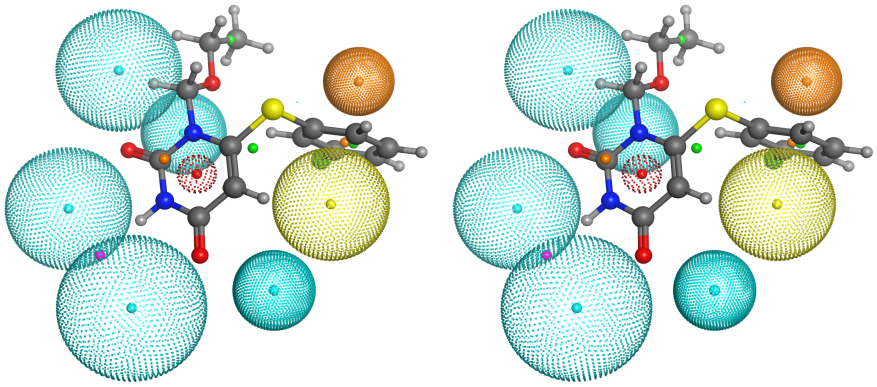


Figura 2.3: *Representació estereogràfica del mapa farmacofòric d'una molècula orgànica. Els farmacòfors colorats de blau indiquen la presència d'acceptors de pont d'hidrogen, les zones caracteritzades per regions que contenen sistemes amb electrons π s'indiquen en taronja i les zones amb volums d'exclusió en groc.*

– Les relacions estructura-activitat (*Structure-Activity Relationships*, SAR). Són mètodes que intenten relacionar el valor d'una determinada activitat amb una o més propietats moleculars. Existeixen nombroses variants d'aquests models segons la naturalesa de l'activitat a predir. Quan aquesta correspon a un valor quantificable, el mètode rep el nom de QSAR (*Quantitative* SAR), en els quals el valor de l'activitat es descriu com una funció matemàtica d'una o més variables, que corresponen a propietats estructurals. El nom pot variar fins i tot segons el paràmetre objectiu: en el cas de tractar-se d'un temps de retenció cromatogràfic (veure secció 7.4, pàg. 184) es parla de relació estructura-retenció quantitativa (*Quantitative Structure-Retention Relationship*, QSRR). De forma genèrica, si l'activitat se substitueix per una propietat en general, es considera una relació quantitativa estructura-propietat (QSPR). Els mètodes QSAR han demostrat ser extremadament útils per a la predicció de propietats d'interès tant per a la química mèdica com per al disseny de fàrmacs.[22–24] Per aquest motiu, han estat els principals mètodes que s'han utilitzat en aquest treball per a l'establiment de models de predicció.

2.1 Mètodes de predicció basats en l'estructura molecular

"*Apporte-moi seulement de quoi lire: des sangliers, des canards, des poulets, de la pâtisserie, de la cervoise... Tu lis dans la cervoise, aussi? Si elle est bien tirée, elle devient très lisible.*" Astérix le Gaulois - Le Devin (1972)

En emprar el mot "predicció" dins de l'àmbit de la quimioinformàtica hom ha de tenir cura d'interpretar-lo degudament, entenent-lo en tot moment com una probabilitat d'esdeveniment basada en raonaments científics, més que un vaticini dictaminat per un ordinador. És necessari fer aquesta

apreciació (per molts òbvia), per tal d'entendre la subtilesa dels resultats obtinguts, ja que depenen en gran mesura de com es defineix, s'aplica i s'interpreta el mètode emprat en cada cas.

Una vegada definit l'objectiu d'un estudi, cal que la informació de què disposem pugui ser traduïda a un llenguatge que pugui ser manipulat per un ordinador. Des del punt de vista químic, una de les dificultats més grans a superar és la descripció dels sistemes moleculars implicats. Això implica tant la seva estructura com les propietats que es deriven d'ella.

2.1.1 Química *in silico*

Els paràmetres disponibles per al tractament computacional de compostos químics reben el nom de **descriptors moleculars**. Cada descriptor permet representar matemàticament una propietat, traduint el seu significat químic a un valor numèric tractable per un sistema informàtic. Aquest pot ser utilitzat a posteriori en diversos tipus d'estudis com la selecció de quimioteques o l'establiment de models de predicció.

Si bé la quimioinformàtica permet tractar un gran nombre de dades, la representació gràfica d'aquestes (o dels resultats que se'n deriven) suposa una important limitació, motiu pel qual no es pot perdre de vista la conveniència de cercar noves i millors formes per facilitar la seva interpretació.

Donat el gran nombre de descriptors existents es fa molt difícil trobar una classificació inequívoca per a tots ells. Per tenir una idea, el programari DRAGON v.6[25] (que normalment es pren de referència per aquest tipus de càlculs) disposa de més de 4000 descriptors. Això condueix a realitzar classificacions segons la conveniència de cada cas.

En l'elaboració d'aquest treball els descriptors que s'han utilitzat es poden classificar en 6 grups, segons la manera com extreuen la informació. Donat que tots ells s'han calculat mitjançant el programari MOE[26] es considera adequat mantenir la nomenclatura dels descriptors utilitzada en ell. Al capítol 8 (pàg. 203) es mostra una classificació diferent, atenent al tipus d'informació que permeten obtenir.

1. **Descriptors basats en la connectivitat:** Cada molècula és representada per una taula on figura cada element amb la seva càrrega i la connectivitat entre sí, indicant l'ordre d'enllaç. D'aquesta manera no depenen de la conformació de adopta el sistema molecular.

    - Propietats físiques. Utilitzant aquesta informació de forma directa es poden calcular propietats com el pes molecular, la càrrega formal de la molècula, la seva densitat o el nombre d'àtoms d'un determinat tipus. Per altra banda, aquesta informació es pot combinar amb models de predicció basats en la contribució atòmica, per obtenir propietats més complexes: la refractivitat molar, el coeficient de partició octanol/aigua (logP) o la solubilitat aquosa, entre molts d'altres.
    - Propietats de mida i forma. Descriptors tals com els índexos de Kier & Hall utilitzen grafs moleculars per estimar diferents aspectes de la forma del sistema. La contribució atòmica de cada àtom sobre la superfície accessible de van der Waals també es pot aproximar a partir de la taula de connectivitat, permetent que descriptors com SlogP_VSA$_i$ ho combinin amb el valor d'una segona propietat atòmica.

- Càrregues parcials. Combinant la informació de la connectivitat amb dades de l'electronegativitat de cada àtom, el mètode PEOE, per exemple, permet estimar el valor de les càrregues parcials.

2. **Descriptors basats en la matriu d'adjacència:** La matriu d'adjacència ($\mathbb{A}$) associada a una molècula amb n àtoms correspon a una matriu $n \times n$. Cada element $\{\mathbb{A}\}_{ij}$ pren el valor de 1 si els àtoms i i j es troben enllaçats; d'altra manera val zero. Si enlloc de tenir en compte només la connectivitat, es té també en compte l'ordre d'enllaç es parla de matriu topològica. Finalment, si se substitueix l'ordre d'enllaç per la distància entre els àtoms i i j, s'obté la matriu de distàncies.

   A més de poder extreure informació de la suma de rengles d'aquestes matrius, com el diàmetre molecular o l'índex de Petitjean, hi ha descriptors que es defineixen a partir de modificacions més complexes de la matriu d'adjacència. Els descriptors BCUT i GCUT, per exemple, es defineixen a partir dels valors propis d'una matriu d'adjacència modificada per un segon tipus de descriptor (típicament valors de càrregues parcials o estimacions del logP).

3. **Descriptors farmacofòrics:** Aquests descriptors assignen a cadascun dels àtoms pesants presents en la molècula una determinada tipologia, dependent d'ell i del seu entorn, que correspon a una característica farmacofòrica. D'aquesta manera cada àtom que no sigui hidrogen (que són eliminats durant el càlcul), pot acabar classificat com a donor, acceptor, àcid, bàsic o hidrofòbic. En basar-se tan sols en el tipus d'àtoms presents (*atom types*), aquests descriptors no depenen de l'estructura tridimensional de la molècula.

4. **Descriptors basats en el càlcul de l'energia:** Donat que l'energia d'un sistema depèn de la seva geometria, aquests descriptors depenen de l'estructura tridimensional de la molècula. L'energia (o qualsevol dels seus components) es pot estimar mitjançant qualsevol dels mètodes exposats al capítol 1. L'aplicació de mètodes semiempírics o SCF permeten utilitzar altres observables com a descriptors (com per exemple el moment dipolar o l'energia dels orbitals HOMO i LUMO).

5. **Descriptors basats en l'estructura:** La informació derivada de la connectivitat estructural o de les càrregues parcials es pot combinar amb la informació conformacional del sistema. D'aquesta manera es pot avaluar l'efecte de l'estructura sobre descriptors. Així, hom és capaç d'avaluar només un component espacial del moment dipolar, o predir propietats farmacocinètiques en les quals és clau tenir en compte l'estructura tridimensional.

6. **Descriptors basats en claus estructurals:** Aquests descriptors, també anomenats *fingerprints*, codifiquen en forma de vector les característiques estructurals (o farmacofòriques) d'una molècula. Cada posició correspon a una característica, que pot ser diferent en funció del *fingerprint* utilitzat. Si la subestructura és present a la molècula, el valor de la posició en questió pren el valor 1. Aquest mètode de codificació permet agilitzar notablement els processos de cerca estructural i de semblança.

2.1.2 Una estructura per una activitat

Encara que els descriptors moleculars no permetin aconseguir una definició tan exhaustiva de l'essència de les molècules com ho fa la mecànica quàntica, sí que permeten descriure'n el comportament. A mitjans del segle XX, Hammett fou el primer a descriure la reactivitat química com la combinació lineal d'un component estèric i un electrostàtic, en estudiar l'efecte del substituent sobre la reactivitat en un conjunt de compostos.[27,28] Tot i que aquest treball fou matisat i ampliat posteriorment per Taft,[29] la idea que una propietat química pogués calcular-se a partir d'altres (més fàcilment mesurables o calculades matemàticament), va obrir el camí a Hansch et al. (1964) a estudiar la correlació entre l'activitat biològica i el logP.[30,31]

A partir d'aquests i de molt altres estudis (l'anàlisi de Free Wilson, per exemple, que permet determinar l'activitat biològica a partir de la presència o absència de determinades subestructures a l'interior de la molècula[32]), s'establí la idea que qualsevol propietat molecular pot ser determinada a partir de la seva estructura.[33]

Si és així, donat que els descriptors moleculars permeten descriure propietats moleculars a partir de l'estructura del compost, es poden utilitzar com a variables independents del model de predicció. És a dir, es pot ajustar una funció matemàtica que, per un conjunt de dades d'entrenament, permeti correlacionar (linealment o no) el valor dels descriptors moleculars amb el valor d'activitat. L'establiment del model QSAR pot realitzar-se, per tant, de diferents maneres segons la manera com es proposi l'equació matemàtica.

Regressions lineals múltiples

L'aproximació més senzilla per establir un model QSAR és considerar que la resposta desitjada (y) es pot obtenir com una combinació lineal d'un o més (n) descriptors (x_i).

En cas de dependre de tan sols una variable, el model es pot obtenir aplicant el mètode de regressió lineal de l'ajust per mínims quadrats. Tanmateix, la situació normal acostuma a ser la contrària, i els models solen dependre de més d'un descriptor. En aquests casos cal utilitzar mètodes de regressió lineal múltiple, com per exemple el mètode *Partial Least Squares* (PLS).

Per tal d'incorporar totes les variables en el model, el mètode PLS intenta minimitzar l'error quadràtic mitjà (*Root-Mean Square Error*, RMSE) cercant la millor combinació lineal de variables ortogonals t_i(eq. 2.1). Aquestes, definides alhora com combinació lineal dels descriptors moleculars, es defineixen de tal manera que permetin explicar no tan sols la variància dels descriptors sinó també la variància dels valors esperats (eq. 2.2).

$$y = c_0 + \sum_{i=1}^{n} c_i t_i \tag{2.1}$$

$$t_i = \sum_{k=1}^{n} c_{ik} \cdot x_k \tag{2.2}$$

El PLS és un dels mètodes més utilitzats en QSAR per a l'establiment de models de predicció de valors quantitatius. D'altra banda, tal com es veu en el capítol 8 (pàg. 203), hom pot desitjar utilitzar

la metodologia QSAR per classificar. En aquests casos, es discretitza el valor de la funció resposta y en tants valors com classes es poden tenir: $\{0,1\}$ en el cas de classificacions binàries (actiu/no actiu, és permeable/no és permeable, etc.) o $\{0,1,2\}$ en el cas de ternàries.

D'altra banda, l'anàlisi discriminant lineal (*Linear Discriminant Analysis*, LDA) permet calcular l'equació de l'hiperplà que separa millor les diferents classes. El seu algorisme d'optimització té com a objectiu maximitzar el nombre de dades ben classificades, cosa que fa de forma anàloga al PLS, considerant la combinació lineal de les variables.

Ambdós mètodes de regressió lineal (PLS i LDA), consideren que la relació entre els descriptors i el valor a predir es pot descriure de forma lineal. Malauradament aquesta afirmació no és certa en tots els casos. Per aquest motiu es proposaren mètodes que acceptessin relacions no lineals entre les variables d'entrada. Des del punt de vista del QSAR destaquen les xarxes neuronals artificials, que foren plantejades com a alternativa al PLS a finals del segle XX,[34] i que actualment li comencen a fer ombra.

Capítol 3

Mètodes basats en la intel·ligència artificial

La curiositat humana ha constituït des de sempre el motor principal per promoure l'estudi de la nostra pròpia naturalesa. D'entre totes les qüestions que mai s'hagin plantejat, la que pren més rellevància en el context d'aquest treball és la referent a l'essència de la nostra intel·ligència. Com podem pretendre definir la intel·ligència artificial (IA), si no definim en primer lloc el que entenem per intel·ligència?

Anaxàgores (550 aC) introduí el concepte de *nous* (intel·ligència) com una substància infinita i immutable, responsable de controlar qualsevol ésser viu. Tot i així, aquesta elegant però provocadora definició primigènia no s'adapta al concepte que acceptem actualment per referir-nos a la intel·ligència. El cert és que el concepte -o el grup de conceptes relacionats- representat per la paraula "intel·ligència" es troba en permanent evolució i s'ha intentat definir (sense èxit) de moltes maneres diferents. Per la consegüent discussió es tindran en compte dues possibles definicions, seguint les idees recollides per R.W.Howard:[35]

La intel·ligència com a característica intrínseca del comportament. Considera que és el comportament d'un individuu el que pot ser intel·ligent o no. Aquesta definició es troba d'acord amb la naturalesa variant respecte de les èpoques del propi concepte d'intel·ligència, donat que el judici pel qual es determina si una acció pot ser considerada intel·ligent o no, depèn de molts factors tals com el marc temporal, les motivacions o el context en què es produeix l'acció. Tot i que el llindar, doncs, és una decisió merament arbitrària, un comportament intel·ligent es troba normalment associat a un comportament adaptatiu. Aquesta apreciació esdevindrà important en la secció 3.2 (pàg. 67), quan es descriguin els algorismes genètics.

La intel·ligència com a conjunt d'habilitats. Dins d'aquest marc, són les habilitats mentals les que determinen si un individuu és intel·ligent o no. El debat sobre quines capacitats mentals (de les múltiples que es poden determinar) han de ser incloses dins d'aquesta definició, continua obert: habilitats sensorials, memòria, capacitat d'adaptació, sensibilitat per a la música, etc. El llistat de possibles habilitats mentals és difícil d'enumerar, i fins i tot pot subdividir-se en habilitats específiques com

la capacitat verbal: per exemple, A.Turing, personatge de principis del s.XX que ha pres una importància més que notable en la teoria de la computació, considerà que la intel·ligència es troba en la capacitat de respondre com un ésser humà.

Així doncs, hom conclou que no existeix una definició única i determinista de la intel·ligència i cal recórrer a definicions consensuades i actualment acceptades.

Encara que comprenguem què volem expressar quan parlem d'intel·ligència, la seva simulació requereix una definició mecanística de la seva naturalesa. En aquest punt apareix una pregunta capital, que ha marcat tot el treball realitzat entorn a la IA: **És possible simular computacionalment la intel·ligència?** És probable que cadascú opini de diferent manera, però de ben segur que totes les respostes troben cabuda en una de les quatre posicions descrites per R.Penrose[a]:[36]

- $\mathscr{A}$) Tot pensament és computació. Conegut com a funcionalisme o IA forta.
- $\mathscr{B}$) El coneixement és un aspecte de l'acció física del cervell. Com tota acció física, aquesta pot ser simulada computacionalment, però la simulació no pot generar coneixement (IA feble).
- $\mathscr{C}$) El coneixement és generat per l'acció física del cervell, i és aquesta la que no pot ser degudament simulada.
- $\mathscr{D}$) El coneixement no es pot explicar en termes físics ni computacionals.

Aquest treball tan sols s'ha pogut dur a terme negant autoritàriament el punt de vista $\mathscr{D}$, donat que un dels objectius principals ha estat la implementació de mètodes que intenten simular l'acció física del cervell i emprar-los per a la resolució de problemes dins de l'àmbit de la química mèdica. També es rebutja l'afirmació pròpia de la IA forta ($\mathscr{A}$), considerant que redueix tota habilitat mental a processos que poden ser descrits matemàticament en forma d'algorismes. No és inherent en l'ésser humà l'afany de trobar l'ànima, la negació de ser mers individus programats?

Tal i com apunta Penrose, la discussió entre $\mathscr{B}$ i $\mathscr{C}$ esdevé més subtil. Per un costat, $\mathscr{B}$ suposa que disposem de la teoria necessària per entendre i simular exactament el comportament del cervell. Tot i que la neurociència computacional avança ràpidament, les actuals tècniques que cerquen la simulació computacional del funcionament del cervell encara presenten certes limitacions. Molts d'aquests estudis es recolzen en models cognitius (amb més o menys realisme biològic) basats en xarxes neuronals artificials, i les anomenades *spiking neurons*.[37] Qui sap si gràcies al projecte Conectoma Humà[38] i al projecte Blue Brain[39] (promoguts pels EEUU) es podrà, d'aquí a uns anys, conèixer amb exactitud el planell de rutes d'aquest òrgan tan complex i així entendre'l (a ell, i conseqüentment a nosaltres mateixos) un xic millor.

Encara que avui en dia sembla plausible descartar directament el punt de vista $\mathscr{B}$, atenent al fet que encara no es disposa de la teoria necessària per simular correctament l'acció física del cervell, el Dr. Aréchiga recordà al 1998, a la revista de la Reial Acadèmia de Medicina de Catalunya, com de ràpid augmenta el coneixement en la neurociència i la confiança que aquest transmet:[40]

"Hoy se acepta como verdad indiscutida que la actividad mental es producto del funcionamento cerebral. La psicología y la psiquiatría reconocen cada vez más explícitamente sus fundamentos neurobiológicos y no se

[a]S'ha respectat la nomenclatura original del treball

antoja remoto el día en que las características más entrañables de la naturaleza humana, como los sentimientos, los pensamientos y las ilusiones, sean expresables en términos de interacciones eléctricas y químicas en redes neuronales."

Així doncs, sembla plausible descartar el punt de vista $\mathscr{B}$ atenent que encara no disposem de la teoria necessària per simular correctament l'acció física del cervell. Per tant, l'acceptació de $\mathscr{C}$ com l'opció més realista en l'actualitat suposa que tot intent que realitzem per simular el comportament intel·ligent dels organismes vius esdevindrà una aproximació. Aquests intents reben el nom de mètodes d'intel·ligència artificial, algun dels quals s'exposen a continuació (concretament aquells que s'han utilitzat en la realització d'aquest treball).

3.1 Xarxes Neuronals Artificials

La major part dels mètodes basats en la IA reben el nom de l'estructura o comportament natural que pretenen imitar. No és sorprenent, doncs, que les xarxes neuronals artificials (*Artificial Neural Networks*, ANN) siguin mètodes matemàtics basats en la neuroanatomia i el funcionament del cervell.

Tal com indiquen Russel i Norvig,[41] l'explicació de les ANN pot entendre's de dues maneres diferents. El primer punt de vista ja s'ha anat comentant anteriorment, i constitueix la voluntat de concebre les ANN com anàlegs matemàtics dels processos que tenen lloc dins del cervell. Aquesta visió contrasta amb un segon punt de vista que apel·la a la naturalesa merament computacional del procés, i on les ANN són considerades un conjunt d'elements de processament aritmètics, disposats en forma de xarxa, que actuen com a funcions mboxbooleanes i l'ús dels quals és vàlid i legítim independentment de qualsevol semblança biològica.[42]

Tanmateix, aquesta disjuntiva no hauria de sorprendre després de l'exposició de l'apartat anterior: els dos punts de vista es poden reconciliar apel·lant a la declaració $\mathscr{C}$ sobre la IA, on la definició computacional pot ser considerada com l'aproximació matemàtica de l'activitat neuronal descrita en el primer punt. Per aquest motiu cal remarcar la importància del mot "*artificial*" en el nom d'aquests mètodes, i no oblidar que són aproximacions computacionals de processos que encara no estem en condició de poder simular correctament. En aquest treball se seguirà el primer punt de vista exposat, tot considerant al segon d'ells la base mecanística sobre la qual se sustenta el primer.

3.1.1 Breus apunts de neurociència

El sistema nerviós està format a grans trets per un conjunt de cèl·lules, altament especialitzades, anomenades neurones. Al cervell, aquestes cèl·lules es troben interconnectades entre sí en forma de xarxes i són les responsables de processar la informació que reben dels òrgans sensorials.

El procés de visió és un clar exemple de la interconnexió neuronal: la llum que entra dins la cavitat ocular i assoleix la retina desencadena una resposta per part de les neurones sensorials que finalitza amb la generació d'un estímul, que s'envia a les terminals nervioses que tenen associades. Aquesta informació es transmet pels nervis (en forma d'un pols elèctric) fins arribar al cervell, on les seves neurones processen la informació i permeten la creació d'una hipòtesis de percepció. Inevitablement cal parlar d'hipòtesi, doncs el que hom veu és la interpretació que fa el seu cervell del canvi conformacional de molècules de retinal localitzades en les proteïnes receptores de la seva retina. Aquest fet és especialment rellevant per exemplificar els límits de la teoria que emprem per a simular el cervell. Quan veiem un gos, no tan sols associem la representació mental que ens fem de l'objecte, sinó que som capaços de classificar-lo dins de la classe dels animals mamífers i fins i tot ens podria despertar records oblidats d'aquell gos que havíem tingut de petits.

Les característiques estructurals de les neurones són claus per permetre la seva relació, i establir així, el canal per on es transfereix i es processa la informació. Les estructures cel·lulars responsables de rebre els estímuls provinents de les neurones veïnes i transformar-los en un senyal químic s'anomenen dendrites (figura 3.1). La naturalesa d'aquest senyal correspon a un gradient electroquímic (que implica el bombeig de cations entorn a la membrana cel·lular de la neurona) i es propaga al llarg de tota la cèl·lula fins arribar a l'axó, on s'envia la informació a la següent neurona. El procés pel qual té lloc la transmissió intraneuronal de la informació rep el nom de sinapsi.

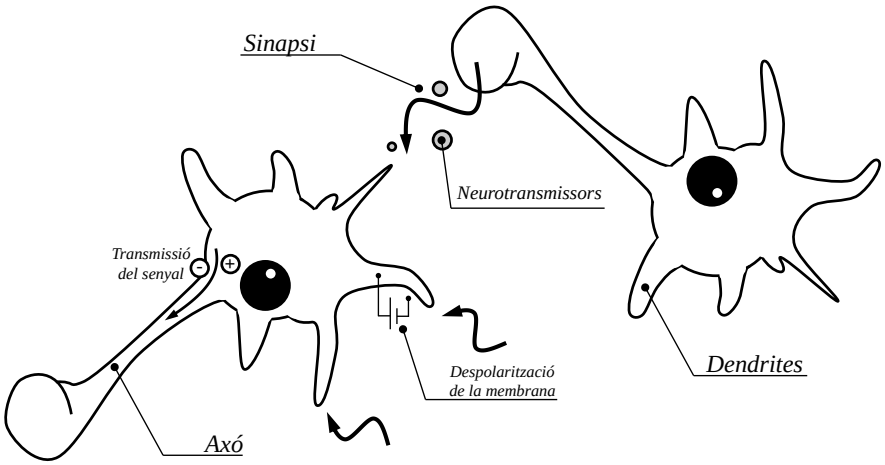


Figura 3.1: *Esquema il·lustratiu de la transmissió de la informació en neurones biològiques.*

Tot i així, no totes les neurones es troben permanentment actives, sinó que són activades quan reben un senyal d'alguna o diverses neurones a les quals es troba associada. El processament i la definició de la senyal de sortida depèn del conjunt de totes les informacions rebudes.

3.1.2 Imitant a la natura

De la descripció que s'ha realitzat anteriorment del comportament de les neurones biològiques, hom pot concloure que aquestes estructures actuen com a centres de processament on es transforma la informació que reben de l'exterior. Així doncs, la unitat bàsica de processament de les xarxes neuronals artificials rep el nom de **neurona artificial**. A partir d'aquest punt, s'assumeix que qualsevol elusió al concepte de neurona fa referència a la seva vessant artificial.

Actualment, el model de simulació neuronal proposat per McCullogh i Pitts és el més utilitzat. Segons aquest model, les neurones es consideren unitats de processament que reben un conjunt de senyals d'entrada, que combinen i transformen abans d'enviar-les com a senyal de sortida a totes les neurones a qui estan connectades, figura 3.2. En cas d'haver-hi només una unitat de processament, aquesta rep el nom de perceptró.

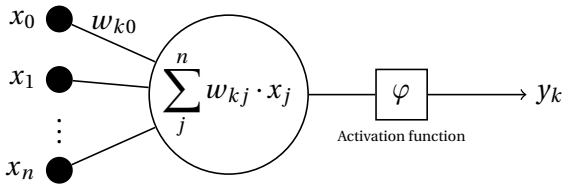


Figura 3.2: *Representació esquemàtica d'un perceptró, on x_i representen els senyals d'entrada i w_{kj} els pesos de cada connexió sinàptica. La informació es combina linealment abans d'aplicar la funció d'activació (φ) per rendir la resposta final (y_k) de la neurona.*

El pes de les neurones

En l'intent d'imitar el cervell, les ANN formen (com el seu nom indica) una xarxa d'unitats de processament que es troben interconnectades entre si. Encara que es puguin definir diferents esquemes, en tots ells se simulen les connexions sinàptiques a partir d'un valor numèric anomenat **pes**. D'aquesta manera, les connexions prenen més rellevància a mesura que el valor del pes que la governa augmenta en valor absolut, enfortint-la en cas que sigui positiu o inhibint-la si és negatiu.

L'estudi de l'evolució dels pesos durant el procés d'entrenament d'una ANN, pot donar una idea de les connexions menys rellevants per al model, ajudant a identificar els senyals d'entrada prescindibles. Aquest procés es coneix amb el nom de *pruning* i és de molta utilitat per a la selecció de descriptors (secció 5.4, pàg. 119).

L'arquitectura neuronal

Normalment els pesos sinàptics s'agrupen en forma de matriu, creant l'anomenada **matriu de pesos**, on l'element w_{kj} correspon al pes de la connexió entre la neurona k i la j. Normalment, no totes les neurones podran estar connectades amb totes, la manera com es connecten les diferents unitats estructurals porta a parlar de l'**arquitectura** de la xarxa neuronal.

Les ANN s'han definit clàssicament en forma de capes.[43] Prenent el nom de *feed-forward networks* o ANN multicapa, les connexions sinàptiques es produeixen entre neurones de diferents capes i de forma unidireccional. Aquest tipus de xarxes gaudeixen d'un ampli recolzament teòric i han demos-

trat ser molt eficaces per a la resolució d'un gran ventall de problemes, entre els quals també s'inclouen referents a la química mèdica.[44–46]

Totes les ANN definides en aquest treball es poden descompondre en tres tipus de capes (figura 3.3):

1. **Capa d'entrada**. Està formada per tantes neurones com de senyals d'entrada (x_i) disposa la xarxa. Les seves neurones esdevenen el punt d'entrada de cadascun dels descriptors del problema.

2. **Capa de sortida**. Conté una o més neurones que, donat un vector d'entrada, retornen el resultat final de la xarxa (o_i).

3. **Capes amagades**. Les neurones situades entre la capa d'entrada i de sortida s'anomenen neurones amagades, que es disposen en forma d'una o més capes. La incorporació d'aquestes neurones fou crucial per resoldre els problemes de forma més eficient i permetre el modelat de sistemes complexos. Tot i que no existeix un mètode determinista per definir el nombre de capes i de neurones amagades, es considera que dues capes amagades és suficient com per garantir la resolució de la gran majoria de problemes.[43]

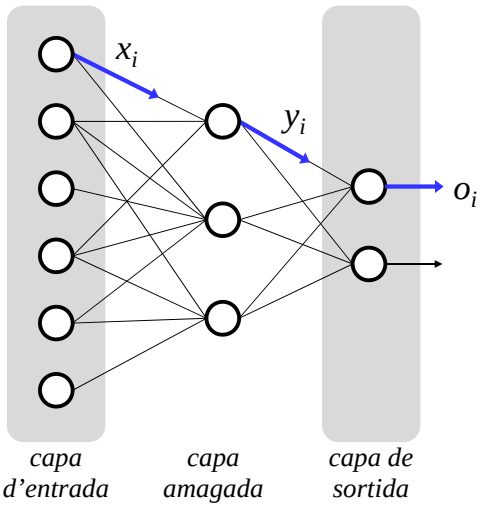


Figura 3.3: *Diagrama esquemàtic d'una xarxa neuronal de topologia 6-3-2.*

El nombre de neurones que conformen cadascuna de les capes defineix la topologia de la xarxa. Si una ANN està formada, per exemple, per 3 neurones d'entrada, dues capes amagades (amb 5 i 4 neurones cadascuna) i dues neurones de sortida, es diu que la seva topologia és 3-5-4-2.

La manera com es relacionen les neurones entre les diferents capes defineix el comportament global de l'ANN.[47] L'arquitectura tradicional que s'utilitza per a les *feed-forward networks* és l'esquema *fully-connected*, que consisteix en connectar cada neurona d'una capa amb totes les neurones de la capa següent.

Actualment es disposa d'altres esquemes que permeten construir xarxes més elaborades: amb arquitectures que disposen de memòria (*feedback networks*)[48] o xarxes tridimensionals.[49]

El tractament del senyal

En primera instància, quan una neurona k rep un conjunt de senyals ($\{x_j\}$), provinents d'aquelles neurones que hi han establert una connexió sinàptica amb un pes no nul ($w_{kj} \neq 0$), actua com un combinador lineal. D'aquesta manera, la neurona incorpora totes les informacions en un sol valor, fent que la importància de cadascun dels senyals sigui proporcional al pes sinàptic de la connexió de la qual provenen:

$$u_k = \sum_j^n w_{kj} \cdot x_j + b_k \tag{3.1}$$

A més dels senyals neuronals, cada neurona de la xarxa (excepte les que conformen la capa d'entrada) reben una entrada addicional anomenada *bias*. Encara que el senyal associat a aquesta entrada es fixa a 1, el seu pes (b_k) es tracta com qualsevol altre. Aquest terme, inclòs dins del tractament lineal del senyal (eq. 3.1), actua com una transformació afí de la resposta neuronal, permetent la translació de la resposta de tal manera que beneficia a l'eficàcia del model global.[50]

Un dels grans avantatges de les ANN per sobre altres mètodes és la seva capacitat d'establir relacions no lineals entre les dades d'entrada i els valors de sortida. Aquesta característica és aportada per la funció d'activació (φ) que, aplicada sobre el resultats de la combinació lineal (u_k), desenvolupa una doble funció: transforma de forma no lineal els senyals d'entrada i limita l'amplitud del resultat (y_k) dins d'un marge d'interès. Es poden definir diferents tipus de funcions d'activació (figura 3.4, pàgina següent) segons la necessitat de cada mètode:

- **Funció esglaó o binària**. Les neurones del model McCulloch-Pitts utilitzen la funció unitària de Heaviside per obtenir una resposta binària entre dos valors $\{0,1\}$. El valor llindar (o *threshold*) s'acostuma a fixar a zero, encara que pot deixar-se a l'elecció de l'usuari. Una funció binària compleix:

$$y_k = \varphi(u_k) = \begin{cases} 0 & \text{si } u_k < 0 \\ 1 & \text{si } u_k \geqslant 0 \end{cases} \tag{3.2}$$

  En casos molt concrets, hi ha mètodes, en què la funció d'activació ha de ser binomial, és a dir que pot prendre només els valors $\{-1,1\}$ (secció 5.2.2, pàg. 106).

- **Funció lineal**. Correspon a una funció definida a trossos, on cadascun dels trams és definit mitjançant una recta. La funció resposta augmenta de 0 a 1 de forma lineal, en un marge $(-a,a)$. Si els valors a processar cauen dins d'aquesta regió, la neurona actua com un combinador lineal.

$$y_k = \varphi(u_k) = \begin{cases} 0 & \text{si } u_k \leqslant -a \\ u_k & \text{si } -a < u_k < a \\ 1 & \text{si } u_k \geqslant a \end{cases} \tag{3.3}$$

- **Funcions contínues**. Es poden definir funcions que siguin contínues i diferenciables en tot el marge d'u_k. Les funcions sigmoïdals són les més utilitzades en xarxes neuronals multicapa, en cas de desitjar una resposta entre $[0,1]$, ja que presenten un bon balanç entre el comportament

lineal i no lineal:

$$y_k = \varphi(u_k) = \frac{1}{1+e^{-\theta \cdot u_k}} \tag{3.4}$$

Per altra banda, si cal que la funció resposta sigui antisimètrica respecte de l'origen, es pot aplicar la funció de la tangent hiperbòlica:

$$y_k = \varphi(u_k) = \tanh(u_k) \tag{3.5}$$

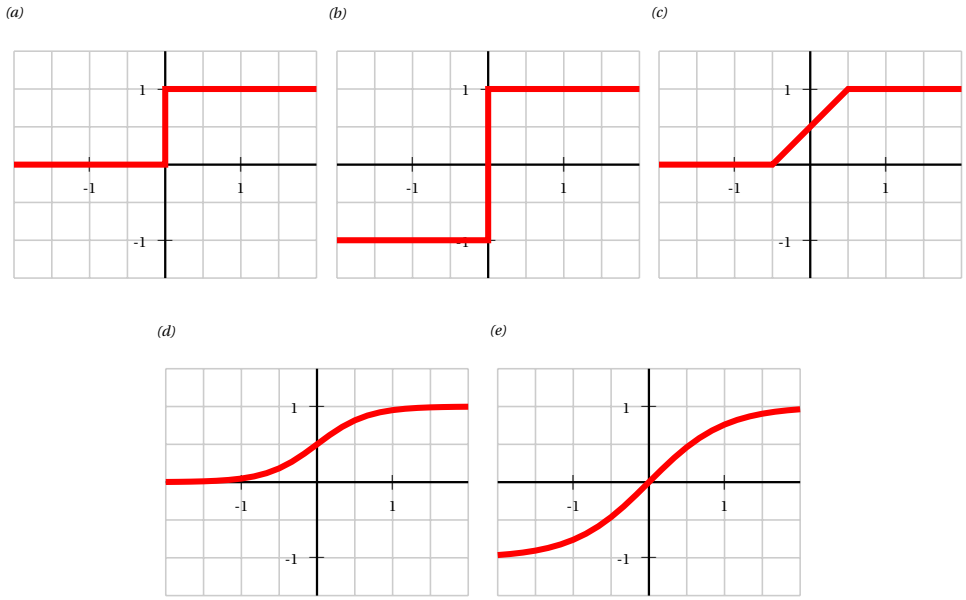


Figura 3.4: *Representació de les principals funcions d'activació utilitzades en el desenvolupament del treball: funció binària (a), binomial (b), lineal amb un paràmetre $a = 0.5$ (c), sigmoïdal (d) i tangent hiperbòlica (e).*

3.1.3 Estratègies per entrenar el coneixement neuronal

Quan es presenta un vector amb p senyals d'entrada, l'ANN els processa segons el que s'ha comentat fins ara, obtenint en últim terme una resposta. Aquesta, segons l'eq. 3.1, serà funció dels valors dels pesos sinàptics que regeixin cadascuna de les connexions presents a la xarxa.

Mètodes d'entrenament supervisat

Els **mètodes d'entrenament supervisat** s'encarreguen de cercar aquella combinació de pesos que fan possible que una ANN pugui predir correctament el major nombre d'entrades. Per aquesta finalitat requereixen d'un conjunt de vectors d'entrada de resultat conegut (conjunt d'entrenament, *training set*), que serveixin per optimitzar els pesos de forma iterativa.

Partint d'un conjunt aleatori, es presenten una o més entrades a la xarxa neuronal. El resultat o resultats obtinguts per cadascun dels vectors es comparen amb els corresponent valors esperats. A continuació, diferents algorismes permeten recalcular els pesos de tal manera que la diferència entre el valor esperat i calculat esdevingui mínima. El procés es repeteix de forma iterativa fins aconseguir l'error mínim. En el cas d'ANN multicapa, en les quals s'hagi definit una funció d'activació sigmoïdal,

el mètode d'optimització més utilitzat és l'algorisme de retropropagació de l'error (veure secció 5.3, pàg.111), que tot i no correspondre's amb el funcionament real del cervell, gaudeix d'un gran suport teòric que l'avala.[50]

Hom pot avaluar la capacitat de predicció del model obtingut amb un segon conjunt de dades (conjunt de validació, *test set*), que no hagi participat en el procés d'entrenament. És molt interessant realitzar aquesta validació interna de forma prèvia a la selecció d'un model, per avaluar la seva capacitat real de predicció (veure secció 4.3, pàg. 85). El millor model serà aquell que, tot i presentar bons resultats durant l'entrenament, conserva la plasticitat suficient com per adaptar-se a un senyal desconegut.

En cas de només tenir en compte el conjunt d'entrenament, es pot donar el cas (relativament comú) que la xarxa s'adapti tan bé a aquestes dades, que perdi la capacitat de realitzar prediccions coherents fora d'aquest entorn. Aquest efecte es coneix amb el nom d'*overfitting*.[51]

L'*overfitting* pot tenir diverses causes, totes elles relacionades amb el fet d'aportar un excés d'informació en l'etapa d'entrenament:

1. El nombre de neurones amagades afecta de forma directa a la capacitat d'adaptació de la xarxa al conjunt d'entrenament. En augmentar el nombre de neurones amagades s'augmenta la capacitat de correlacionar les dades d'entrada amb la resposta esperada, tant que pot arribar a significar l'ajust de l'error experimental. Davant d'aquesta situació, el conjunt de validació es troba amb una xarxa molt especialitzada que no el sap tractar correctament. Com que no existeix el mètode per determinar inequívocament el nombre de capes i neurones amagades[52] tots els estudis realitzats en aquest treball comencen per una optimització de la topologia de l'ANN.

2. Seguint amb la idea del punt 1, l'especialització de la xarxa augmenta amb el temps d'entrenament. Per aquest motiu, una de les maneres que es proposa per reduir l'efecte de l'*overfitting* és disminuir el nombre d'iteracions de l'entrenament[53] (apartat 5.1.1, pàg. 103).

3. Si realment l'ANN és capaç d'ajustar l'error de les dades, pot significar que rep un excés d'informació per part de les neurones d'entrada. Així doncs, l'*overfitting* pot ser indicatiu de la necessitat de reduir el nombre de descriptors utilitzats per al model QSAR, que es pot realitzar mitjançant un mètode de *pruning*[54] (secció 5.4, pàg. 119) o mitjançant l'optimització dels descriptors utilitzats amb algorismes genètics[55] (secció 5.5, pàg. 121).

Mètodes d'entrenament no supervisat

Quan no es desitja utilitzar (o no es disposa d'informació referent) les respostes de cada conjunt d'entrada, els mètodes d'entrenament no supervisat permeten identificar característiques comunes entre vectors d'entrada i agrupar-los. Els *Self-Organizing Maps* (SOM) o xarxes de Kohonen[56] són un d'aquests mètodes i el seu ús s'ha estès ràpidament dins de l'àmbit de la quimioinformàtica[57] i la química mèdica.[58,59]

A diferència de les xarxes neuronals multicapa, els SOM permeten preservar la topologia, és a dir, les relacions entre les neurones d'entrada són preservades per les neurones de sortida.[60] Com a

resultat s'obté un mapa bidimensional on s'ha projectat l'espai n-dimensional d'entrada. Per aquest motiu, els SOM poden utilitzar-se per a l'estudi de la semblança entre un conjunt de dades descrites en espais multidimensionals o avaluar les agrupacions que formen.

La disposició de les neurones en els SOM també és diferent, essent potser la que més s'assembla als sistemes biològics. Les neurones es disposen sobre un espai bidimensional $n \times m$, definit per l'usuari (figura 3.5). Totes les neurones reben tots els components del vector d'entrada, i normalment els pesos que governen aquestes connexions es recullen en forma de columna, donant als SOM forma de cub.

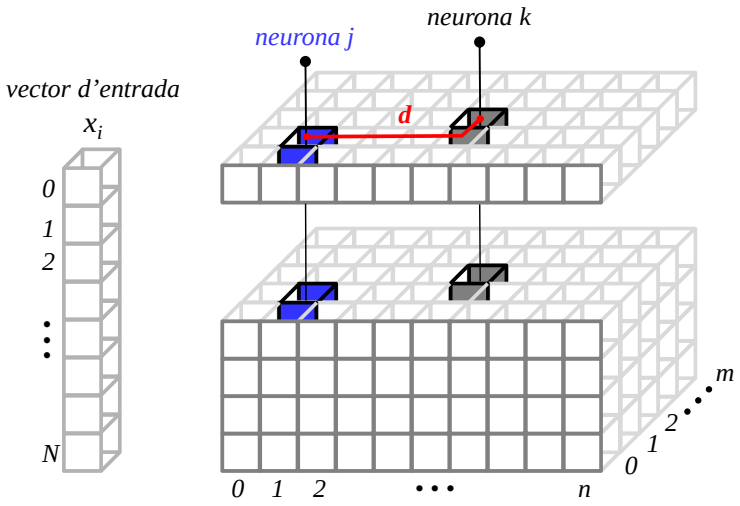


Figura 3.5: *Representació esquemàtica d'un SOM o xarxa de Kohonen.*

Durant el procés d'entrenament, les neurones es veuen obligades a competir entre elles per decidir quina s'estimularà. Aquesta competència es resol a través d'un mètode iteratiu, on cada iteració per cadascun dels vectors d'entrada segueix l'algorisme 1.[50,61]

Algorisme 1 Mètode d'aprenentatge no supervisat en SOM.

1: Inicialització aleatòria dels pesos sinàptics (w)
2: Presentació d'un vector d'entrada x_i a la xarxa
3: Selecció de la neurona k, amb pesos més semblants a x_i segons l'eq. 3.6
4: Actualització dels pesos de la neurona k (w_{ki}) i del seu entorn més proper (eq. 3.7)

La identitat de la neurona k guanyadora es determina segons la distància euclidiana entre el vector d'entrada x_i i els seus pesos w_{ki}:

$$k \leftarrow \min\{\sum_{j=1}^{N}\left(x_j - w_{kj}\right)^2\} \tag{3.6}$$

A continuació, per a l'actualització dels pesos d'una neurona j, es té en compte una funció de la

distància ($f(d)$) que separa aquesta neurona de k, fent que l'efecte de l'actualització disminueixi a mesura que d augmenta.

$$w_{ji}(t) = w_{ji}(t-1) + \eta(t) \cdot f(d) \cdot \Big(x_i - w_{ji}(t-1)\Big) \tag{3.7}$$

3.2 Algorismes genètics

Els algorismes genètics (*Genetic Algorithms*, GA) són un mètode d'optimització que codifica les variables a tractar de tal manera que poden ser considerades com els gens constituents del cromosoma d'un organisme, que segueixen el procés d'evolució natural. A través de la selecció natural, la reproducció sexual i la mutació, els cromosomes van mostrant en cada nova generació millor adaptació, quantificada matemàticament en una funció objectiu (*fitness*) corresponent a la funció a optimitzar.

Els inicis dels algorismes genètics es troben a finals dels anys 1950, quan H. J. Bremermann, en el seu treball *The evolution of intelligence*, plantejà una connexió entre l'aprenentatge individual i l'aprenentatge evolutiu. En aquest procés s'utilitzava un model primitiu del que uns anys després acabaria sent l'esquema bàsic dels algorismes genètics.[62,63] Entre 1960 i 1970, a la Universitat de Michigan, John Holland va desenvolupar el concepte d'algorisme genètic pròpiament dit.[64–66] Aquest, inclòs dins del marc de la computació evolutiva (creada en la dècada anterior), va ser concebut amb la intenció de no ser específic per a un tipus de problema, sinó d'adaptar-se al problema a resoldre.[67]

El fet que un algorisme proporcioni uns resultats segons l'adaptació de les solucions a les condicions del problema, s'aproxima al fenomen pel qual un sistema biològic respon a un estímul segons la seva adaptació a l'entorn. Dins l'algorisme cada solució equival a un individu, representat pel seu codi genètic en forma de vector (cromosoma). Al llarg del temps i mitjançant l'ús de diferents operadors de selecció i encreuament s'imiten les condicions d'aparellament i de reproducció que permeten l'evolució de les generacions a unes de millor adaptades a l'entorn.

Aquesta tècnica d'optimització s'ha utilitzat al llarg d'aquest treball per a l'optimització de formulacions cosmètiques (secció 5.5, pàg. 121) i com a alternativa al mètode PLS (pàg. 55) en projeccions sobre el pla $\mathscr{H}^2$ (secció 6.4, pàg. 133).

3.2.1 Nomenclatura bàsica dels algorismes genètics

La terminologia emprada en els algorismes genètics constitueix una barreja de nomenclatura biològica i artificial que deixa palesa l'analogia entre la naturalesa matemàtica del mètode i els sistemes biològics. Els GA treballen sobre cadenes d'informació assimilables a les cadenes que emmagatzemen la informació genètica dels organismes vius en forma de **cromosomes**. Per aquest motiu, cada cadena d'informació s'anomena cromosoma o individu.

Cada cadena conté informació de totes les variables i tots els paràmetres, la variació dels quals determina la resposta. Els cromosomes biològics estan formats per l'agrupació de gens, cadascun dels quals aporta una informació diferent. De la mateixa manera, a cada part en què se segmenta lògicament el cromosoma per desar-hi la informació d'un paràmetre se l'anomena **gen**. Aquests gens

poden prendre diferents valors, anomenats **al·lels**. El tipus i el nombre d'al·lels disponibles venen determinats per la manera en què es codifica la informació a l'interior dels cromosomes, figura 3.6. D'altra banda, el conjunt de tots els cromosomes d'una població rep el nom de genotip. Quan s'expressa el cromosoma i el genotip interacciona amb el seu entorn i es descodifica, es parla de fenotip.

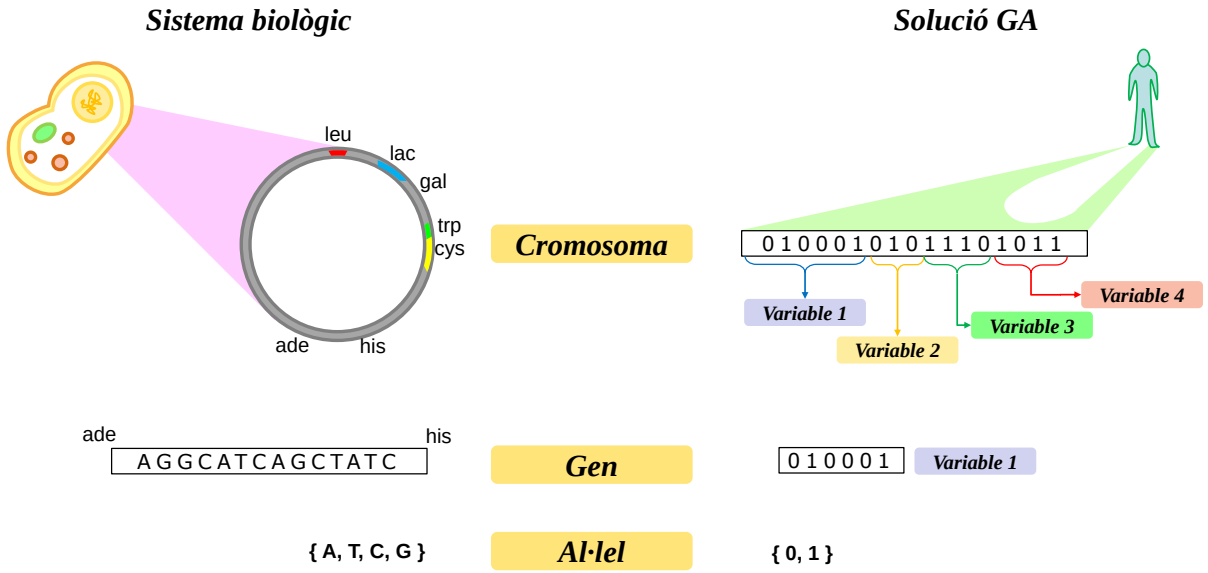


Figura 3.6: *Semblança entre la nomenclatura biològica i la pròpia dels GA.*

Dins l'algorisme, als individus millor adaptats en el medi els correspon una major probabilitat de reproduir-se. La relació entre aquesta i la probabilitat mitjana de selecció de tots els cromosomes s'anomena pressió de selecció.

3.2.2 L'algorisme genètic al descobert

La figura 3.7 representa com evoluciona un conjunt de solucions al llarg del procés evolutiu. A continuació es comenten les principals característiques de cadascuna de les etapes.

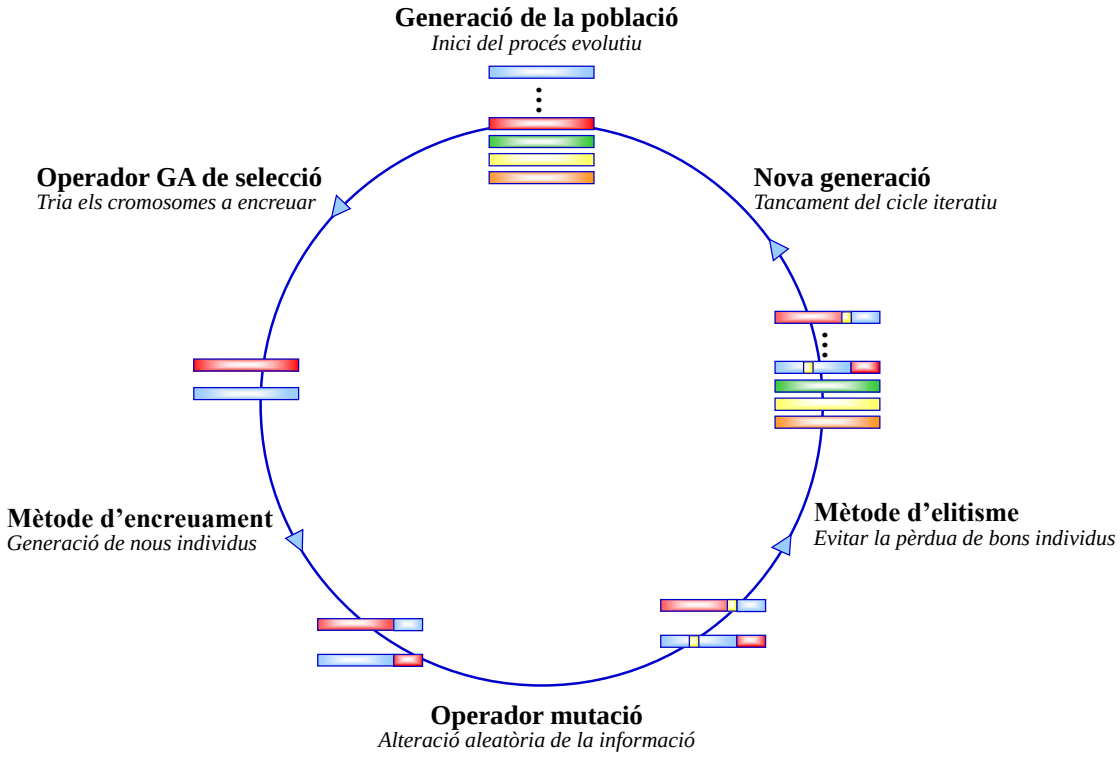


Figura 3.7: *Representació esquemàtica de l'evolució d'una població a través de GA.*

Codificació de les dades

Per tal que els cromosomes puguin actuar, cal traduir el problema a una funció matemàtica avaluable (*fitness*) i dependent d'uns certs paràmetres (gens), que solen ser descriptors moleculars. La codificació del cromosoma més utilitzada és la codificació binària, en la qual el cromosoma està format per una seqüència d'uns i zeros (en aquest cas només es disposa de dos al·lels), on cada bit representa la presència (1) o absència (0) d'un paràmetre. Aquesta codificació s'ha emprat en l'optimització del nombre de descriptors moleculars necessaris per a la descripció de formulacions cosmètiques (secció 5.5, pàg. 121).

En cas de ser necessaris més de dos al·lels s'acostumen a utilitzar cromosomes de nombres enters. Tanmateix, aquests han demostrat presentar problemes de convergència quan l'espai de resposta ve definit per un gran nombre de descriptors.[68]

Operador GA de selecció

L'operador GA de selecció és l'encarregat d'iniciar el procés evolutiu. En aquest procés es determina quina parella d'individus s'escullen per a la reproducció. Tot el procés està regit per una **probabilitat de reproducció** que determina la capacitat dels cromosomes d'ésser seleccionats, essent major per aquells individus amb major valor de *fitness*. D'aquesta manera s'imposa una certa restricció als cromosomes de la població relacionable amb el procés de selecció natural. Amb l'aplicació d'aquest operador s'inicia el procés de reproducció que inclou, a més de la pròpia selecció, els operadors d'encreuament i mutació.

La selecció es realitza successivament fins completar la mida de la població. Cal tenir en compte que un determinat nombre de cromosomes pares, normalment una parella, crearan el mateix nombre de cromosomes fills. Hi ha una gran varietat d'operadors GA de selecció, entre ells:

- ***Roulette Wheel Selection***. És un dels operadors més senzills i es basa en la imitació del funcionament d'una ruleta. De la mateixa manera que en el joc d'atzar es reparteix l'espai segons el nombre de possibles eleccions, cada posició representa un cromosoma que pot ser escollit. Aleshores, es genera un nombre aleatori (que representa la bola que gira) que determina el cromosoma seleccionat segons el marge en què quedi comprès.

  L'espai es pot repartir equitativament per tots els cromosomes o definint una probabilitat de reproducció (P_i), fent que l'espai ocupat per cada cromosoma sigui proporcional al valor del seu *fitness* (f_i):

$$P_i = \frac{f_i}{\sum_j f_j} \tag{3.8}$$

- ***Stochastic Sampling Without Replacement***. És una variació del mètode anterior en què es pretén millorar el manteniment de l'heterogeneïtat de la població, impedint la selecció preferent dels individus amb millor *fitness*. Igual que en l'operador anterior, l'espai es reparteix segons la probabilitat de reproducció de cadascun dels cromosomes. En aquest cas, emperò, el nombre aleatori que determina la probabilitat de selecció del primer element és funció de mida de la població que es vol seleccionar: si l'objectiu és seleccionar n cromosomes, es genera un nombre aleatori entre 0 i $1/n$. Els següents individus se seleccionen a partir d'aquest punt situant-se en intervals de $1/n$ unitats.

- ***Stochastic Remainder Selection Without Replacement***. Els cromosomes seleccionats per a la reproducció i el nombre de còpies que es realitza de cadascun d'ells ve determinat pel l'equació:

$$e_i = P_i \cdot N \tag{3.9}$$

  on e_i correspon al nombre de còpies esperades de l'individu i, P_i a la probabilitat de reproducció (eq. 3.8) i N a la mida de la població.

- ***Tournament Selection*** Es trien a l'atzar un nombre q de cromosomes per formar un subconjunt de la població i se'n selecciona el millor. Aquest procediment es realitza tants cops (μ) com

individus calgui seleccionar per completar la generació següent. Habitualment $q = 2$, cas en què es parla de *tournament* binari.

La probabilitat de reproducció d'un individu determinat (P_i), correspon a:

$$P_i = \frac{1}{\mu^q}\left((\mu - i + 1)^q - (\mu - i)^q\right) \tag{3.10}$$

Encreuament

La recombinació genètica, que permet explorar nous punts de l'espai de solucions, se simula computacionalment encreuant la informació genètica de dos cromosomes parentals. En la imitació del procés biològic, la **taxa d'encreuament** en determina la probabilitat.

La majoria dels mètodes d'encreuament (o *crossover*) es dissenyen segons la necessitat de cada programador. Cas de no restringir-se el nombre i la naturalesa dels al·lels (limitacions en el nombre d'elements amb un valor determinat o en valors que no es poden repetir), els mètodes generals solen ser:

- **Encreuament en un punt**. Es defineix un punt de tall que indica una posició a partir de la qual els cromosomes parentals s'intercanvien per crear els dos cromosomes fills. Tot i ser el mètode d'encreuament més intuïtiu, la seva principal limitació és que no permet obtenir totes les combinacions possibles dels gens que formen els cromosomes pares, de manera que es restringeix l'espai de respostes explorat.

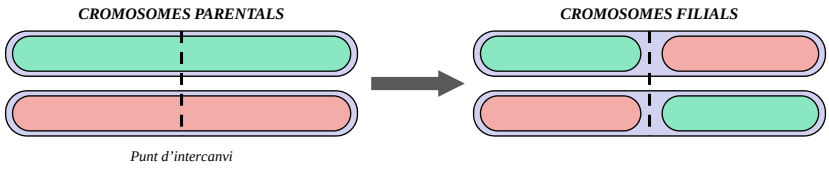


Figura 3.8: *Esquema de l'encreuament en un punt. A partir del punt de tall definit a l'atzar s'intercanvien les cadenes d'informació.*

- **Encreuament en dos punts**. Mètode anàleg a l'anterior que introdueix un segon punt de tall, fent que l'inici i el final d'un cromosoma fill prové d'un mateix pare. Aquesta incorporació permet ampliar parcialment l'espai de respostes explorat.

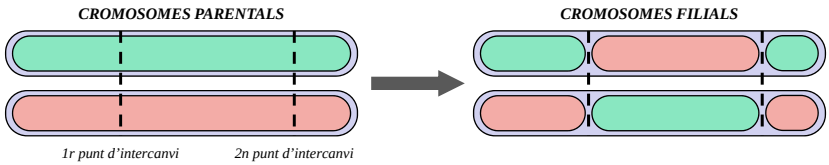


Figura 3.9: *Esquema de l'encreuament en dos punts. S'intercanvia la informació dels cromosomes en l'interval comprès entre els dos punts de tall definits aleatòriament.*

- **Encreuament uniforme**. La millora introduïda per l'encreuament en dos punts condueix a la consideració de mètodes amb un gran nombre de punts de tall. Així, en l'encreuament uni-

forme, cada posició que forma la cadena del cromosoma fill es copia d'un pare determinat a l'atzar. Tot i que l'espai de resposta que pot explorar-se és molt gran, amb aquest mètode es poden perdre bons individus a causa de l'alteració massiva de la població.

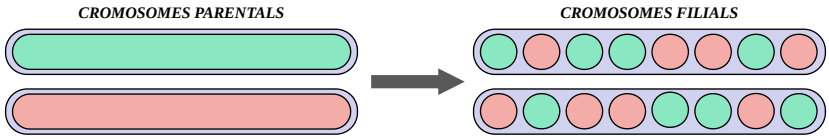


Figura 3.10: *Esquema de l'encreuament uniforme. Cada gen filial té la mateixa probabilitat de provenir d'un o altre cromosoma pare.*

Mutació

Els organismes vius són susceptibles a canvis aleatoris en la seva estructura interna. Alguns, com en el cas dels *Single Nucleotide Polymorphisms* (SNP), consisteixen en canvis en la seqüència genètica i són responsables de les diferències individuals.[69] Aquests canvis, que reben el nom de mutacions, poden ser deguts a diferents factors, tals com errors en la replicació del material genètic o agents externs.

En l'intent de simular el comportament dels sistemes biològics, els GA han d'ésser proveïts d'un component aleatori que simuli aquestes mutacions. En l'algorisme, el procés de mutació no és més que l'alteració d'una posició del cromosoma a l'atzar, amb una probabilitat anomenada **taxa de mutació**, que contribueix a mantenir la diversitat de la població.

Elitisme

Mentre té lloc el procés de reproducció, les operacions d'encreuament i de mutació poden provocar alteracions als individus amb major *fitness* i la seva conseqüent pèrdua. Aquest fet es tradueix en problemes de convergència que alhora incrementen notablement el temps de càlcul. Per tal d'evitar-ho, existeixen mètodes que afavoreixen els millors individus de la població incrementant la seva probabilitat de sobreviure i passar a la següent generació sense veure's alterats.

Tanmateix, hom ha de ser conscient que un abús d'aquestes tècniques pot induir a problemes d'homogeneïtzació de la població (tots els individus acabarien sent còpies del mateix cromosoma).

Amb la finalitat d'evitar la pèrdua dels millors individus, De Jong introduí l'any 1975 la tècnica de l'elitisme,[70] que consisteix en el pas automàtic dels millors individus de la població parental a la nova generació. Així, es prevenen els canvis en els al·lels dels millors cromosomes i s'assegura la seva selecció. L'elitisme no implica que els millors individus de la població parental no es puguin creuar, ans el contrari: els pares es copien directament a la generació filial però participen també en el procés de reproducció.

La selecció del nombre d'individus que passen directament a la nova generació pot fer-se en base a:

- La transferència directa d'un determinat nombre d'individus de la població inicial a la nova.

- La creació d'una població virtual intermèdia (sense tenir en compte l'elitisme) i posterior selecció dels millors individus (ja siguin de la generació inicial o intermèdia). En aquest cas existeix la possibilitat que un cromosoma dominant (amb bon *fitness*) de la generació intermèdia sigui idèntic a un pare i, per tant, que la generació filial contingui individus repetits.
- La tècnica equivalent a l'anterior, permetent tan sols la selecció sense repeticions dels millors individus.

Objectius

Amb tot el que s'ha exposat en els capítols anteriors, es proposen per aquesta tesi doctoral els següents objectius:

1. Desenvolupament d'un programari per a la implementació de mètodes basats en intel·ligència artificial, particularment xarxes neuronals artificials, i la seva adaptació per ser aplicat a problemes de la química mèdica.

2. Validació del programari desenvolupat contrastant els resultats obtinguts amb estudis publicats de referència. Aplicació dels mètodes implementats en el disseny de fotosensibilitzadors per a la teràpia fotodinàmica del càncer, concretament en la predicció de propietats fisicoquímiques com el màxim d'absorció i el caràcter hidrofòbic. Igualment, ús del programari per a la predicció de la localització subcel·lular de fotosensibilitzadors tetrapirròlics.

3. Estudi de noves tècniques de representació bidimensional d'espais multidimensionals, que permetin facilitar la interpretació dels resultats computacionals obtinguts.

Part II

Artifical Intelligence Suite (ArIS)

El programari *Artificial Intelligence Suite* (ArIS) s'ha desenvolupat en aquesta tesi íntegrament al laboratori de Disseny Molecular del GEM, al departament de Química Orgànica de l'IQS, amb la finalitat d'ajudar a resoldre problemes relacionats amb la química mèdica.

Aquesta part es divideix en tres capítols, el primer dels quals (capítol 4) introdueix l'arquitectura interna del programari. ArIS s'estructura de forma modular atenent al gran ventall d'aplicacions en les quals pot utilitzar-se. Això suposa un gran avantatge ja que permet incloure fàcilment noves eines de càlcul sense modificar el cos principal del programari.

Els mètodes de càlcul de què disposa ArIS es descriuen en el capítol 5. Entre ells destaca l'algorisme de retropropagació de l'error, àmpliament utilitzat per a l'entrenament de xarxes neuronals artificials multicapa en l'establiment de models de predicció no lineals.

Finalment a *Reflexions sobre la representació gràfica dels resultats*, es descriuen alguns dels aspectes gràfics com la interfície disponible per l'usuari i la representació d'espais multidimensionals sobre un pla, que s'ha utilitzat en capítols posteriors (part III).

Capítol 4

Implementació general d'ArIS

Els models QSAR són actualment una de les eines LBDD més emprades en quimioinformàtica. Una de les principals limitacions dels programaris que s'utilitzen per establir aquests models és la manca de mètodes no lineals. Per suplir aquesta mancança i aprofundir en el coneixement dels algorismes, es proposa la creació d'un programari que incorpori mètodes d'intel·ligència artificial.

D'altra banda, en els darrers anys s'ha estat desenvolupant al laboratori de Disseny Molecular del GEM el programari PRALINS (*Program for Rational Analysis in Silico*)[71] per a la selecció racional de quimioteques combinatòries, que s'ha utilitzat àmpliament amb bons resultats.[72] Així doncs, no es pot perdre de vista que la creació d'un nou programari sigui necessàriament compatible amb el ja consolidat al grup.

Aquesta premissa condiciona els fonaments estructurals del nou programari: el llenguatge de programació. Per optimitzar la compatibilitat entre programaris s'hereta el llenguatge utilitzat en el desenvolupament de PRALINS (llenguatge C de programació). Aquest llenguatge (amb les seves variants C++ i C#) és un dels més utilitzats actualment, desbancant a Fortran o Pascal. A més, el seu coneixement pot facilitar el posterior aprenentatge d'altres llenguatges com Java, Python o Perl, que actualment estan prenent força en l'àmbit de la quimioinformàtica. Així doncs, les discussions que es plantegen en les següents seccions es troben dins del marc del llenguatge de programació C i empren els comandaments i les estructures de control pròpies d'aquest llenguatge.

4.1 Sistema d'arxius i emmagatzematge d'informació

ArIS es presenta en tres parts, a cadascuna de les quals correspon una carpeta d'arxius: el codi font (`src`), les llibreries (`include`) i l'arxiu executable (`bin`).

L'estructura interna del codi font d'ArIS es troba dividida alhora en diferents mòduls, cadascun dels quals té assignada una tasca concreta, i on la interconnexió entre ells defineix el programa final. A més de permetre un major control del flux d'informació global que es genera en l'aplicació (facilitant la depuració del programa i la identificació de possibles errors de programació), aquest disseny augmenta l'eficiència de l'aplicació, definint zones de codi compartides entre els diferents mètodes de càlcul. D'aquesta manera, l'estructura global del programari ArIS segueix l'algorisme 2.

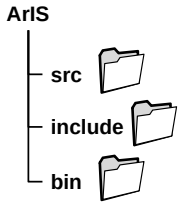


Figura 4.1: *Diagrama del sistema d'arxius d'ArIS.*

Algorisme 2 Flux principal del programari ArIS.

Requeriments: Arxiu d'entrada complimentat

1: `main.c`: Definició del tipus de càlcul i inicialització de les variables globals
2: `1a_read.c`: Lectura de les dades d'entrada
3: `1b_error.c`: Previsió d'errors
4: `2_run.c`: Execució del càlcul definit per l'usuari
5: `5_write.c`: Escriptura dels resultats

Cada mòdul està format a la vegada per un conjunt de funcions i subrutines que es complementen i col·laboren en la realització d'una determinada tasca. Així, l'execució del càlcul per part del mòdul `2_run.c` comporta la crida a funcions especialitzades d'altres mòduls auxiliars (com l'arxiu `3_learning.c` per l'entrenament supervisat d'ANN o `9_GA.c` per l'optimització amb algorismes genètics).

Definició de les variables internes

Tal com s'ha vist a la secció 3.1 (pàg. 59), la quantitat d'informació que es genera i es processa durant l'entrenament i l'aplicació de les xarxes neuronals artificials no és en absolut menyspreable. Per aquest motiu l'emmagatzemament de la informació ha de ser prou efectiu com per garantir un ràpid accés a la informació. Gràcies a l'estructura de punters que ofereix el llenguatge de programació C, gran part de l'emmagatzemament de la informació es realitza definint matrius de dades. Les variables comunes, disponibles des de qualsevol mòdul de càlcul, reben el nom de **variables globals** i es defineixen a la llibreria `global.h`. A continuació es detalla el significat i la codificació emprada per les variables globals.

- **Topologia de la xarxa**. La informació referent a la topologia de la xarxa s'emmagatzema dins el vector `topology`. Aquest vector determina tant el nombre de capes amagades (n_{hidden}) com el nombre de neurones presents en cadascuna de les capes que conformen la xarxa.

  La primera posició del vector informa del nombre de capes amagades de què disposa la xarxa, mentre que la resta de posicions contenen el nombre de neurones presents en la capa d'entrada, en les capes amagades i en la capa de sortida, per aquest ordre. La reserva de la primera posició i les dues posicions corresponents a les capes d'entrada i sortida fa que aquest vector tingui `topology[0]`+3 posicions (de 0 fins a `topology[0]`+2), figura 4.2.

Figura 4.2: *Representació esquemàtica de l'estructura de la variable* `topology`*, per a una xarxa amb quatre neurones d'entrada, dues capes amagades (amb dues i tres neurones respectivament) i una neurona de sortida*

- **Nombre total de neurones**. El nombre total de neurones que conformen la xarxa ($\mathcal{N}$) es calcula a partir de la informació emmagatzemada en la variable `topology`:

$$\mathcal{N} = \sum_{i=1}^{\texttt{topology[0]+2}} \texttt{topology}[i] \tag{4.1}$$

- **Matriu de resultats neuronals**. Aquesta matriu conté els valors numèrics de sortida de cadascuna de les neurones que formen la xarxa neuronal. La implementació del programari ArIS ha considerat dos mètodes possibles per a l'emmagatzemament dels resultats, a fi d'aconseguir la màxima eficiència:

Mètode I:
Emmagatzemament de la informació en una matriu i un vector: la matriu conté les respostes de totes les neurones de la xarxa i el vector conté els resultats esperats de cada entrada.

Mètode II:
Emmagatzemament de la informació en dues matrius: una amb la informació definida per l'usuari a l'arxiu d'entrada, i l'altra amb la resposta de cadascuna de les neurones de les capes amagades i de la capa de sortida.

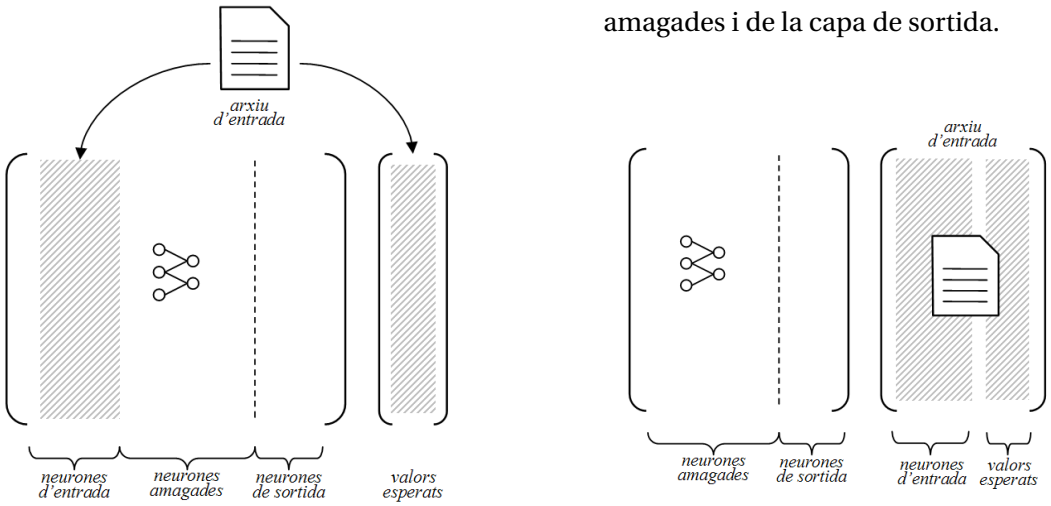


Figura 4.3: *Models considerats inicialment per a l'emmagatzemament de les variables d'entrada.*

Tot i que la implementació del mètode I suposa una substancial complexitat en el procés de lectura de les dades d'entrada (respecte al mètode II), permet agilitzar l'accés a l'informació durant el procés d'entrenament, reduint el temps de càlcul. A més, és conceptualment més còmode, donat que es disposa de la informació de sortida de totes les neurones de la ANN en

una sola variable i només cal accedir al vector de respostes quan es realitza un entrenament supervisat.

D'aquesta manera, el resultat de cada neurona es recull en la matriu `net`. Les columnes d'aquesta matriu corresponen a les neurones de la xarxa, i les fileres a cadascuna de les entrades definides en l'arxiu d'entrada, figura 4.4.

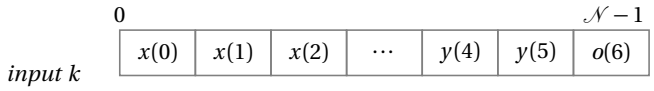


Figura 4.4: *Representació esquemàtica del vector* `net` *associat a un input determinat. Els valors d'entrada s'indiquen com x(i), el resultat de les neurones amagades com y(i). El valor de sortida global de la xarxa és o(6), que en aquest cas correspon a una única neurona de sortida.*

- **Matriu de connexions sinàptiques**. La matriu `synapse` ($\mathbb{S} \in \mu(\mathcal{N} \times \mathcal{N})$) conté la informació relativa a l'arquitectura de la xarxa o la manera com les neurones estan connectades entre si. El valor de l'element s_{ij} determina si les neurones i i j es troben connectades ($s_{ij} = 1$) o no ($s_{ij} = 0$). Per defecte, la variable s'inicialitza segons l'arquitectura *fully connected* en què totes les neurones d'una capa estan connectades amb les neurones de la capa següent. En cas de desitjar una distribució diferent, cal definir-la dins l'arxiu d'entrada mitjançant el comandament `#synapse`.

$$\mathbb{S} = \begin{pmatrix} s_{00} & s_{01} & \cdots & s_{0,\mathcal{N}-1} \\ s_{10} & s_{11} & \cdots & s_{1,\mathcal{N}-1} \\ \vdots & \vdots & \ddots & \vdots \\ s_{\mathcal{N}-1,0} & s_{\mathcal{N}-1,1} & \cdots & s_{\mathcal{N}-1,\mathcal{N}-1} \end{pmatrix}$$

- **Matriu de pesos**. Cada connexió present en la xarxa neuronal té associat un pes (*weight*) que determina la seva importància dins del conjunt. El seu signe dictamina si la connexió és excitativa o inhibitòria i és, essencialment, la informació que s'optimitza durant el procés d'entrenament. Per defecte, cada element de la variable `weight` ($\mathbb{W} \in \mu(\mathcal{N} \times \mathcal{N})$) s'inicialitza aleatòriament en l'interval $[-0.5, 0.5]$. És possible definir altres valors inicials seleccionant l'opció '`Load initial weight matrix`' dins de l'arxiu d'entrada, indicant el nom de l'arxiu que conté els pesos.

  Aquesta informació s'emmagatzema, en un primer moment, en una matriu temporal ($\mathbb{W}'$) que és independent de l'arquitectura de la xarxa. Amb aquest esquema s'obre la possibilitat d'implementar xarxes que puguin variar la seva arquitectura amb el temps (apartat 5.4, pag. 119). Donat que l'establiment d'una connexió sinàptica entre neurones depèn de la matriu `synapse`, la matriu de pesos final (que s'utilitza dins del cos del programa) queda totalment definida amb el producte de Hadamard entre la matriu temporal de pesos i la matriu de connexions sinàptiques:

$$\mathbb{W} = \mathbb{W}' \odot \mathbb{S} \tag{4.2}$$

- **Vector de *bias***. El valor de *bias* de cadascuna de les neurones de les capes amagades i de sortida es disposa en forma de vector, formant la variable `bias` (b). Donat que aquest valor està associat a una neurona i no a una connexió, i que el nombre de neurones d'entrada ve determinat per `topology[1]`, la dimensió de b correspon a $\mathcal{N}$-`topology[1]`.

  Tot i que, a la pràctica, els valors de *bias* es tracten com un pes més, la inicialització dels valors per defecte d'aquesta variable no es troba limitada entre $[-0.5, 0.5]$.

4.2 Escalat de les dades

La informació que es facilita al mètode de càlcul pot ser molt heterogènia, de manera que l'aplicació d'un escalat ajuda considerablement a l'obtenció de resultats acceptables. Per aquesta finalitat, s'implementen a ArIS dos mètodes diferents d'escalat, que l'usuari pot seleccionar mitjançant l'arxiu d'entrada.

1. **Escalat per interval**. Mitjançant aquest mètode totes les variables queden acotades entre 0 i 1, permetent el tractament de components amb qualsevol escala de valors. És un mètode molt sensible a la presència de punts anòmals i permet identificar-los fàcilment, ja que se situen sobre els eixos o allunyats de la resta de punts.

$$x'_i(j) = \frac{x_i(j) - x_{min}(j)}{x_{max}(j) - x_{min}(j)} \tag{4.3}$$

2. **Autoescalat**. Aquest mètode consisteix en una transformació z-dimensional en la qual l'origen de coordenades se situa en el centroide del núvol de punts. El resultat s'expressa en unitats de la desviació estàndard dels punts (σ), tal com es pot veure en l'equació 4.4.

$$x'_i(j) = \frac{x_i(j) - \overline{x}_j}{\sigma_j} \tag{4.4}$$

4.3 Mètodes de validació interna

Per tal d'avaluar la capacitat predictiva dels models durant el procés d'entrenament, les bases de dades utilitzades pel seu establiment es divideixen normalment en un conjunt d'entrenament pròpiament dit (*training set*) i un conjunt de validació interna (*validation set*). El primer s'utilitza per a l'aprenentatge del model i els seus resultats són sovint indicatius de la bondat del model. Amb el conjunt de validació interna es verifica la capacitat de predicció que té el model per aquelles dades que no s'han utilitzat durant l'etapa d'entrenament.[73] Per aquest motiu, es considera necessari introduir mètodes interns de validació als algorismes d'aprenentatge, que puguin ajudar a l'elecció del millor model, figura 4.5.

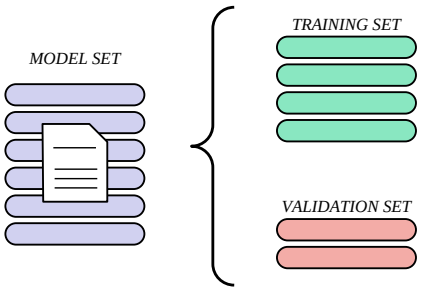


Figura 4.5: *Representació esquemàtica de la divisió d'una base de dades durant el procés d'entrenament per incloure la validació interna.*

Tots els mètodes de validació interna implementats en ArIS es basen en la declaració d'un conjunt de validació, que es reserva durant el procés d'entrenament i es mostra a posteriori al model, permetent extreure'n un error associat a la capacitat de predicció. La diferència entre els mètodes disponibles recau en la manera com es realitza la divisió de la base de dades global (*model set*) i en l'estratègia que se segueix per presentar els dos conjunts al model.

ArIS inclou els mètodes *split-sample validation*, *split-half cross-validation*, *leave-one-out cross-validation* i *k-fold cross-validation* que es descriuen a continuació. La selecció del model a emprar s'especifica en l'arxiu d'entrada amb la paraula clau `#Validation method`, que pot prendre els valors que es detallen en la taula 4.1.

Taula 4.1: *Llistat dels identificadors per a cada mètode de validació interna disponible en ArIS.*

#Validation	Mètode
0	Sense validació
1	*Split-sample validation*
2	*Split-half cross-validation*
3	*Leave-One-Out cross-validation*
4	*K-fold cross-validation*

4.3.1 Split-sample validation

Seguint amb la idea comentada anteriorment, es separa un conjunt de validació del *model set*, que no serà emprat durant l'etapa d'entrenament, figura 4.6. En aquest cas cal definir a l'arxiu d'entrada el nombre d'individus a utilitzar.

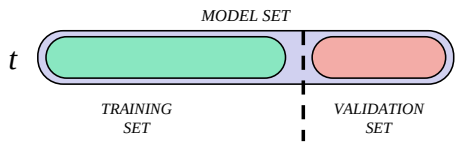


Figura 4.6: *Representació esquemàtica del mètode split-sample validation.*

A l'arxiu d'entrada s'especifica el nombre d'individus que passen a formar part del *validation set* mitjançant l'opció `#Validation_set`. Per defecte aquests han de correspondre a les darreres entrades de la base de dades especificada, cas que es desitgi emprar un conjunt aleatori, cal seleccionar la casella de l'opció `#Use random inputs in valid`.

La identitat del *training set* i del *validation set* es manté constant en tot moment. Això significa que només una part de les dades s'utilitza per entrenar el model.

Tot i ser un mètode que presenta una alta variància, es recomana per a la validació interna de models classificadors.

4.3.2 Split-half cross-validation

Tal com indica el seu nom, en aquest tipus de validació s'utilitza la meitat del *model set* per realitzar la validació interna. A diferència del cas anterior, totes les dades s'empren per a l'aprenentatge, encara que no totes al mateix temps. L'estratègia de la validació encreuada (*cross-validation*) es basa en intercanviar la identitat del *training* i el *validation set* en una segona iteració, figura 4.7. L'error de predicció associat al model es calcula aleshores com la mitjana d'ambdues iteracions.

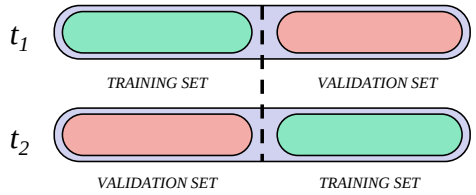


Figura 4.7: *Representació esquemàtica del mètode split-half validation en dues iteracions (t_1 i t_2).*

En aquest cas la variància pot millorar-se amb un major nombre d'iteracions i calculant la mitjana aritmètica de tots els resultats.

4.3.3 Leave-One-Out cross-validation

El mètode de validació *Leave-One-Out* (LOO) és el que permet entrenar amb un major nombre d'individus, ja que només en reserva un per a la validació interna. De la mateixa manera que l'estratègia *split-half*, tots els elements del *model set* formen part del conjunt d'entrenament en un moment o altre, ja que l'entrenament es realitza tantes vegades com individus té el *model set* i cada vegada se'n pren un de diferent per la validació, figura 4.8.

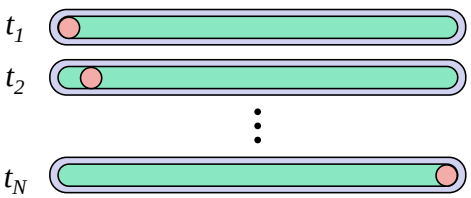


Figura 4.8: *Representació esquemàtica del mètode Leave-One-Out validation, considerant que hi ha N individus en el model set. [• training set, • validation set]*

L'error de la predicció es calcula en aquest cas com la mitjana aritmètica dels errors obtinguts en cada iteració. Un dels principals problemes de la validació LOO és l'aparició d'*overfitting* en grans bases de dades (apartat 3.1.3, pàg. 64).

4.3.4 K-fold cross-validation

Aquest mètode de validació és semblant al LOO, però separa un grup de p elements (en comptes d'un sol individuu) en cada iteració. Per tal que el model pugui veure tots els individus durant l'entrenament, aquesta etapa es repeteix k vegades ($p \cdot k = N$, sent N el nombre d'elements del *model set*), fent que en cada repetició es reservi un grup d'elements diferent per a la validació interna, figura 4.9. De nou, l'error del model es calcula com la mitjana aritmètica de l'error comès en les prediccions, després de cada procés d'entrenament.

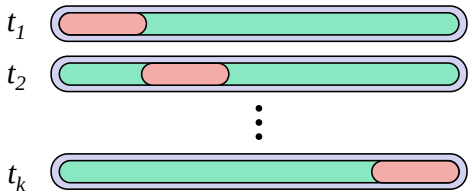


Figura 4.9: *Representació esquemàtica del mètode k-fold validation.[• : training set, • : validation set]*

Hom pot adonar-se que aquest esquema és una generalització dels dos anteriors. Així, LOO correspon a una validació creuada *k-fold* amb $p = 1$ i *split-half* amb $p = N/2$. Per a l'ús d'aquesta validació interna cal definir a l'arxiu d'entrada d'ArIS el nombre d'individus que es desitja emprar utilitzant el comandament `# Validation_set`.

4.4 Mesures d'eficiència i elecció de models

El component aleatori inherent als algorismes d'ArIS fa que la repetició d'un mateix càlcul pugui conduir a resultats diferents. L'establiment d'un model de predicció comporta, doncs, diferents repeticions de l'entrenament. L'avaluació dels resultats obtinguts permet escollir el millor model (fixat per la matriu de pesos i el vector *bias*), l'eficiència del qual es mesura amb criteris estadístics, els quals poden ser aplicats en el conjunt d'entrenament o en el de validació (cas que es defineixi).

Els criteris que s'han emprat en els mètodes de classificació binaris i en els mètodes quantitatius de predicció que es descriuen en els capítols següents són:

- **Mètodes de classificació binaris**:

  1. Nombre d'individus del conjunt d'entrenament classificats correctament, quan l'assignació de la classe dominant es realitza considerant el marge (0.9,1.0] i [0,0.1) per a la classe descartada (*i.e.* $threshold = 0.1$).
  2. Nombre d'individus del conjunt d'entrenament classificats correctament definint un valor de $threshold = 0.25$.

3. Nombre d'individus del conjunt d'entrenament classificats correctament definint com a valor de *threshold* = 0.4.
4. Nombre de falsos positius amb *threshold* = 0.1.
5. Nombre de falsos negatius amb *threshold* = 0.1.
6. Nombre d'individus del conjunt d'entrenament indecisos, on l'assignació de la classe és al 50:50.
7. Error quadràtic mitjà (*Root Mean Square Error*, RMSE) del conjunt d'entrenament.

$$RMSE = \frac{1}{\mathcal{N}} \sum_{i}^{\mathcal{N}} (x_i - y_i)^2 \tag{4.5}$$

8. Paràmetre indicatiu de l'homogeneïtzació dels resultats a una única classe.

9-16. Paràmetres equivalents als anteriors, però definits sobre el conjunt de validació externa.

- **Mètodes quantitatius de predicció**:

1. Nombre d'individus del conjunt d'entrenament predits correctament, prenent com a criteri d'acceptació un error relatiu màxim del 10% (*i.e.* $e = 10\%$).

$$|x_i - y_i| < 0.1 \cdot x_i \tag{4.6}$$

2. Nombre d'individus del conjunt d'entrenament predits correctament amb $e = 20\%$.
3. Nombre d'individus del conjunt d'entrenament predits correctament amb $e = 30\%$.
4. Error quadràtic mitjà de les prediccions del conjunt d'entrenament (eq. 4.5).
5. Màxim dels errors absoluts comesos en el conjunt d'entrenament.

$$E(\text{abs})_{max} = \max_i \{\text{abs}(x_i - y_i)\} \tag{4.7}$$

6. Coeficient de correlació de Pearson (r). Informa de la qualitat de l'ajust obtingut per als individus del conjunt d'entrenament.[74]

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(x_i - \bar{x})^2}} \tag{4.8}$$

7. Paràmetre indicatiu de l'homogeneïtzació dels resultats. Un valor diferent a zero indica que més d'un terç de les prediccions realitzades sobre el conjunt d'entrenament prenen el mateix valor numèric (acceptant una diferència de 0.05 entre el valor predit i l'experimental). Hom pot avaluar amb aquest paràmetre si el model prioritza els valors més abundants, indicant una manca d'adaptació.

8. Mitjana aritmètica dels errors absoluts ($\bar{E}$).

9. Paràmetre d'asimetria (As). Permet estimar (pel conjunt d'entrenament) si la distribució dels errors absoluts (E_i) es troba desplaçada respecte a la seva mitjana, eq. 4.9 i figura 4.10.

$$As = \frac{\frac{1}{N}\sum_{i}^{N}(E_i - \bar{E})^3}{\left(\frac{\sum_{i}^{N}(E_i - \bar{E})^2}{N}\right)^{3/2}} \tag{4.9}$$

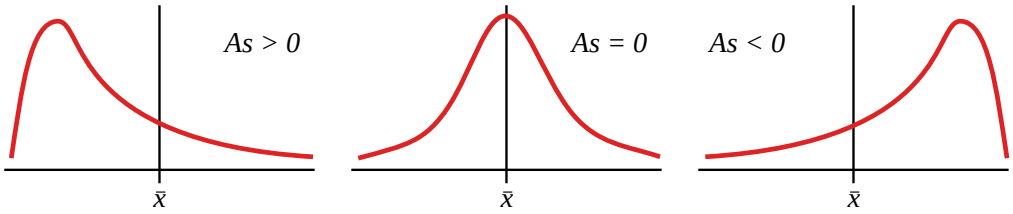


Figura 4.10: *Interpretació gràfica del paràmetre d'asimetria.*

10. Curtosis (C). Paràmetre indicador de la distribució dels errors absoluts del conjunt d'entrenament, eq. 4.10 i figura 4.11.

$$C = \left(\frac{\sum_{i}^{N}\frac{(E_i-\bar{E})^4}{N}}{\left(\sum_{i}^{N}\frac{(E_i-\bar{E})^2}{N}\right)^2}\right)^3 \tag{4.10}$$

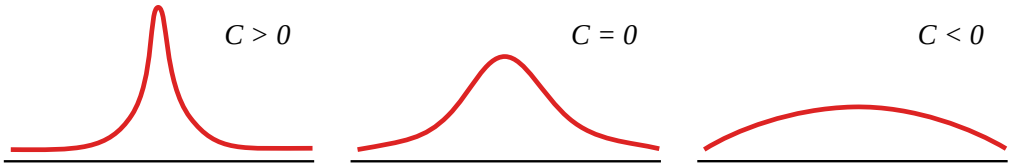


Figura 4.11: *Representació gràfica del paràmetre de curtosis.*

11-20. Paràmetres equivalents als anteriors (1-10), definits sobre el conjunt de validació externa.

21. *Cross-validated correlation coefficient* (q^2). Estima la capacitat de predicció del model

utilitzant, en cas d'haver utilitzat una validació del tipus LOO.[75]

$$q^2 = 1 - \frac{\sum_i^N (y_i - x_i)^2}{\sum_i^N (x_i - \bar{x})^2} \tag{4.11}$$

4.5 L'arxiu d'entrada d'ArIS

4.5.1 Compilació i execució del programari

La compilació del programari ArIS s'ha realitzat mitjançant el compilador gcc4.3-39.1. Per realitzar-ho, en la carpeta corresponent al codi font (`src`) s'inclou el fitxer *makefile*, que conté tots els vincles entre els mòduls de codi i les llibreries necessàries per garantir la generació de la versió executable del programari.

S'ha comprovat el correcte funcionament de l'arxiu *makefile* en diferents distribucions del sistema operatiu linux. Cal, però prendre la precaució de comprovar que la definició de les llibreries matemàtiques pot realitzar-se mitjançant la drecera `-lm`, en cas contrari cal editar l'arxiu anterior i canviar aquest comandament per l'accés a la llibreria `libm.a` corresponent (típicament `/usr/lib/libm.a`).

La compilació es realitza amb el comandament `make`, des de la *shell* d'una terminal linux. Considerant que el camí d'accés a la carpeta d'ArIS es troba definit per la variable `$ARISHOME`, el comandament per compilar el programa correspon a l'expressió següent:

```
:/$ARISHOME/src> make -f Makefile
```

D'aquesta manera es genera l'arxiu executable del programa dins la carpeta `$ARISHOME/bin`. L'execució d'ArIS des d'una terminal s'inicia amb un panell de controls que mostra les diferents opcions de càlcul de què disposa el programari, tant per l'establiment de models com per avaluar-los.

```
************************************************************
           Artificial Intelligence Suite (ArIS)
************************************************************
   Choose an option:
      1.- Single perceptron
      2.- Adaptive Linear Combiner (ADALINE)
      3.- Multi - Adaptive Linear Combiner (MADALINE)
      4.- Multilayer perceptron (Backpropagation Sequential)
      5.- Multilayer perceptron (Backpropagation Batch)
      6.- Input selection via Genetic Algorithms (GNN)
      7.- Self Organizing Maps (SOM)
      0.- Test a model
************************************************************
```

Una vegada l'usuari selecciona un mode de càlcul (corresponent a l'indicador numèric), el programa li demana que indiqui el nom de l'arxiu d'entrada (que conté tota la informació del problema a resoldre). Per tal de facilitar l'*scripting* d'ArIS, s'implementa una drecera que evita la interacció amb l'usuari per a la definició del mode de càlcul i l'arxiu d'entrada. L'usuari pot fer-ne ús afegint darrere el nom del fitxer executable l'identificador numèric del mode de càlcul <*id*> i el nom de l'arxiu d'entrada <*input file*>:

```
:/> $ARISHOME/bin/ArIS.exe <id> <input file>
```

Si es desitja utilitzar, per exemple, un model del tipus perceptró per a la classificació d'un conjunt de dades d'entrada definides en l'arxiu `aris.in`, el càlcul es defineix per mitjà dels comandaments següents:

```
:/> $ARISHOME/bin/ArIS.exe 1 aris.in
```

4.5.2 Definició de l'arxiu d'entrada

L'arxiu d'entrada conté tota la informació que requereix el programari per realitzar un determinat tipus de càlcul. Les opcions generals es recullen en la taula 4.2, i són indispensables per tots els mètodes. Recullen la informació topològica i estructural de la xarxa així com els valors dels vectors d'entrada, utilitzats per l'entrenament i la validació del model.

Taula 4.2: *Opcions generals per al fitxer d'entrada del programari ArIS.*

Comandament	Descripció
`#Title:`	Títol identificador del càlcul realitzat
`#Inputs:`♦	Nombre de neurones d'entrada. Els valors d'entrada es defineixen a continuació, amb el comandament `#InFILE:`, en forma de columnes, separades per un espai i reservant les últimes posicions pels vectors de sortida
`#Output:`♦	Nombre de neurones de sortida. Ha de ser necessàriament 1 en el cas de mètodes de predicció quantitatius i en classificacions amb perceptró o ADALINE
`#Hidden:`	Nombre de capes amagades. El nombre de neurones amagades en cadascuna d'elles cal definir-les amb el comandament `number*:`
`#Threshold:`	Valor llindar per a funcions d'activació
`#Learn rate:`	Valor del *learn rate* (η) per a mètodes d'entrenament basats en la regla δ
`#Trainset:`♦	Nombre d'exemples utilitzats com a conjunt d'entrenament
`#Class:`♦	Nombre de classes possibles
`#OutNumber:`♦	Nombre de columnes definides com a valors de sortida
`#ActivFunction:`♦	Tipus de funció d'ona a utilitzar (pàg. 63)
`#Iddle:`	Màxim nombre d'iteracions a realitzar en cas que el resultat no millori en dues etapes consecutives
`#Method:`♦	Selecció de l'objectiu del model: classificació o predicció
`#Testset:`	Nombre d'entrades utilitzades en la validació externa. Els vectors d'entrada són introduïts després del comandament `#TestFILE:`
`#Validation:`	Definició de mètodes de validació interna. En cas de necessitar indicar el nombre d'entrades a utilitzar, cal definir-les amb l'opció `#Validation_set:` (pàg. 85)

♦: camp requerit.

El mètode de selecció de descriptors mitjançant GA (GNN), requereix d'un conjunt addicional d'opcions, per tal de poder definir els paràmetres necessaris pel procés evolutiu, taula 4.3.

Taula 4.3: *Opcions del mètode GNN per a la selecció de descriptors.*

Comandament	Descripció
`#Number of descriptors to select:`	Nombre de descriptors totals d'entre els quals se'n desitja seleccionar tants com indica l'opció `#Input:`
`#Population size:`	Nombre de cromosomes que conformen una població
`#Selection method:`	Tipus de selecció a emprar. Es troben disponibles els mètodes *Roulette Wheel* (`RW`), *Stochastic Remainder Selection Without Replacement* (`SRSWR`) i *Roulette Wheel* amb rank proporcional (`RWR`)
`#Crossover:`	Mètode d'encreuament utilitzat. S'hi troben implementats els mètodes basats en 1 i 2 punts d'encreuament. La taxa d'encreuament s'indica amb l'addició de la paraula clau `prob.*:`
`#Mutation probability:`	Taxa de mutació. Cal tenir en compte que s'aplica dins del procés d'encreuament
`#Elistism definition:`	Mètode d'elitisme emprat. En el cas de l'elitisme simple, el nombre de cromosomes que passen directament a formar part de la generació filial es defineix amb la paraula clau `prob.*:`
`#gnn_mode`	Selecció del mètode d'aprenentatge desitjat per l'establiment del model amb ANN

A més dels paràmetres generals, ArIS disposa d'un conjunt d'opcions addicionals, que permeten personalitzar el mètode d'aprenentatge, taula 4.4.

Taula 4.4: *Opcions addicionals de càlcul disponibles en ArIS.*

Comandament	Descripció
`#Particular input iterations`	Imprimeix el resultat de cadascuna de les etapes iteratives
`#Scale input values`	Es realitza un escalat lineal dels valors dels descriptors com a pas previ al càlcul.
`#Time dependent learning rate`	Definició d'un η dinàmic, els valor del qual va disminuint progressivament en el temps, evitant la pèrdua de bones solucions.
`#Acyclically pattern present.`	Els vectors d'entrenament es presenten al model de forma aleatòria.
`#Use random inputs in valid.`	Els vectors definits en una validació interna del tipus *split-sample* són seleccionats aleatòriament.
`#Inhibit large calc. control`	Inhabilita el control de convergència. ArIS disposa d'un control intern de la convergència que identifica els processos divergents. Automàticament proposa i aplica una sèrie de mesures i torna a repetir el càlcul. En cas de no millorar, l'atura definitivament.
`#Robustness analysis`	Es realitza un nombre determinat de repeticions del càlcul i ArIS retorna com a resultat el millor d'ells.

Finalment es disposen d'un conjunt d'opcions que fan referència a la lectura d'informació externa al programa, taula 4.5

Taula 4.5: *Opcions de lectura de dades en ArIS. En cas d'ésser activades cal definir el nom del fitxer d'entrada mitjançant la paraula clau* `name:`

Comandament	Descripció
`#Load initial weight matrix`	Lectura dels pesos inicials d'un fitxer d'entrada
`#Set initial architecture`	Definició de les connexions sinàptiques a partir d'un arxiu d'entrada
`#Test_architecture`	Arquitectura utilitzada en l'avaluació externa d'un model. El valor `auto` indica l'ús d'una xarxa *fully-connected*
`#Test_weight matrix`	Lectura dels pesos a utilitzar per a l'avaluació externa d'un model. El valor `auto` indica que s'utilitzi el resultat obtingut de l'entrenament realitzat com a pas previ
`#Test_scale parameters`	Lectura dels coeficients de l'escalat de les dades d'entrada per a l'avaluació externa d'un model. El valor `auto` indica que s'utilitzi el resultat obtingut de l'entrenament realitzat com a pas previ

4.5.3 Exemple d'arxiu d'entrada

```
#Title:  ANÀLEGS_AZT(BP-Batch)              nom identificador del càlcul
CARACTERÍSTIQUES TOPOLÒGIQUES
=============================================================================
#Hidden:  2 neurons*:  6 10              nombre de capes i neurones amagades
#Inputs:  6              nombre de neurones d'entrada
#Output:  2              nombre de neurones de sortida
PARÀMETRES ESTRUCTURALS
=============================================================================
#Learn rate:  0.1
#Trainset:  10             mida del conjunt d'entrenament
#Iddle:  10000             nombre màxim d'iteracions
#ActivFunction:  2              selecció de la funció d'activació
DEFINICIÓ DEL PROBLEMA
=============================================================================
#Method:  Classification [x] | Prediction [ ]
#Class:  2             nombre de classes
#OutNumber:  1             nombre de sortides que es defineixen
SECCIÓ D'ENTRADA
=============================================================================
#InFILE:
```

```
 0.0195   0.94   0.0653   0.01292   14.20   7.1   1
 0.0195   0.94   0.0644   0.01288   14.23   7.1   1
 0.0189   0.99   0.0631   0.01182   14.24   6.7   1
 0.0189   0.99   0.0641   0.01186   14.20   6.7   1
 0.0189   0.99   0.0631   0.01182   14.24   6.7   1
 0.0000   1.07   0.0208   0.01301   19.77   6.4   1
 0.0000   0.99   0.0208   0.01772   16.71   5.5   1
 0.0156   0.98   0.0548   0.01778   17.07   5.5   0
 0.0152   1.03   0.0544   0.01617   17.52   5.8   0
 0.0000   1.11   0.0070   0.01229   19.35   6.8   0
VALIDACIÓ INTERNA DEL MODEL
========================================================================
#Validation method:  1              defineix el tipus de validació interna
#Validation_set:  6              nombre d'entrades assignades al conjunt de validació interna
AVALUACIÓ DEL MODEL
========================================================================
#Testset:  5             mida del conjunt de validació externa
#TestFILE:
 0.0156   0.98   0.0293   0.01785   16.99   5.5   1
 0.0144   1.06   0.0634   0.01462   26.05   6.1   1
 0.0172   1.12   0.0636   0.00923   55.07   7.4   0
 0.0172   1.11   0.0702   0.01063   45.11   7.1   0
 0.0169   1.09   0.0975   0.01189   39.19   6.9   0
INFORMACIÓ ADDICIONAL
========================================================================
 #Particular Input Iterations    [ ]
 #Scale Input Values             [ ]
 #Time dependent learning rate   [ ]
 #Load initial weight matrix     [ ]
 #Acyclically pattern present.   [ ]
 #Test_weight matrix             [x]   name:  auto
 #Test_scale parameters          [ ]
 #Test_architecture              [x]   name:  auto
 #Use random inputs in valid.    [ ]
```

4.5.4 Interfície gràfica

Per tal de facilitat la utilització d'ArIS i contribuir a la seva difusió entre els membres del grups de recerca, es crea una interfície gràfica del programari (ArISv). Aquesta permet tant definir les opcions i les dades de l'arxiu d'entrada com analitzar els resultats obtinguts i organitzar-los segons les necessitats de l'usuari.

ArISv s'ha desenvolupat mitjançant el llenguatge de programació Visual Basic dins de Visual Studio 2010.[76] Els elements de la interfície gràfica es relacionen amb el programari ArIS, que actúa com un mòdul de càlcul extern. Tot i que ArIS s'implementà inicialment en l'entorn linux, s'ha desenvolupat una versió operativa per Windows (compilada mitjançant el programari Visual Studio 2010[76]) que fa possible la seva compatibilitat amb la part gràfica.

Segons la filosofia del nou programa, els càlculs realitzats sobre un mateix conjunt de dades s'agrupen en projectes. La creació d'un projecte es realitza per mitjà de la barra d'eines, figura 4.12.

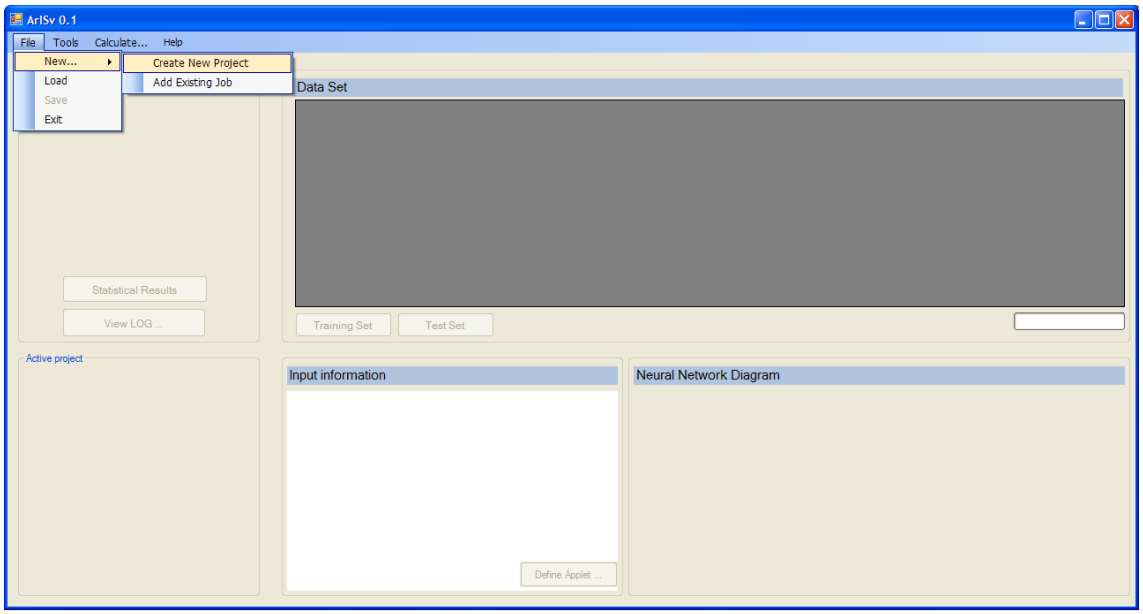


Figura 4.12: *Captura de pantalla de la interfície gràfica d'ArIS: Inici de la sessió.*

El panell contextual que es deriva d'escollir l'opció *Create New Project...* permet definir l'origen i el format de la quimioteca utilitzada per a l'establiment del model. S'incorpora la possibilitat de definir com a arxiu d'entrada llibreries en format *sd-file*. El conjunt de validació externa es pot vincular a un arxiu independent o utilitzar les últimes entrades del conjunt d'entrenament, figura4.13. Finalment la creació d'un nou projecte es finalitza amb la definició d'un nom identificador i el seu lloc d'emmagatzemament.

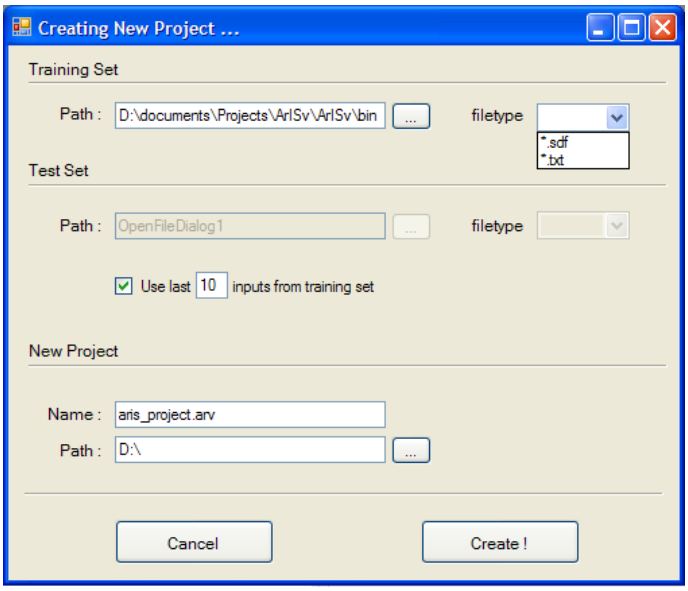


Figura 4.13: *Captura de pantalla de la interfície gràfica d'ArIS: creació d'un projecte nou.*

La base de dades especificada es mostra en forma de taula en la pantalla principal del programa, agrupada en el conjunt d'entrenament (*training set*) i el conjunt de validació externa (*test set*). Es

pot canviar de l'un a l'altre mitjançant els botons *Training Set* i *Test Set*. Els membres de cadascun d'aquests conjunts es poden visualitzar en la part inferior. Per això cal definir el format amb què es desitja fer-ho; els formats suportats en l'actualitat són l'*sd-file* per a la representació gràfica de molècules i formats d'imatge (com jpeg o gif) per als casos més genèrics, figura 4.14. La representació gràfica de molècules es realitza mitjançant el programari lliure Jmol,[77] que permet la visualització interactiva de les molècules podent, fins i tot ,calcular algunes de les seves propietats.

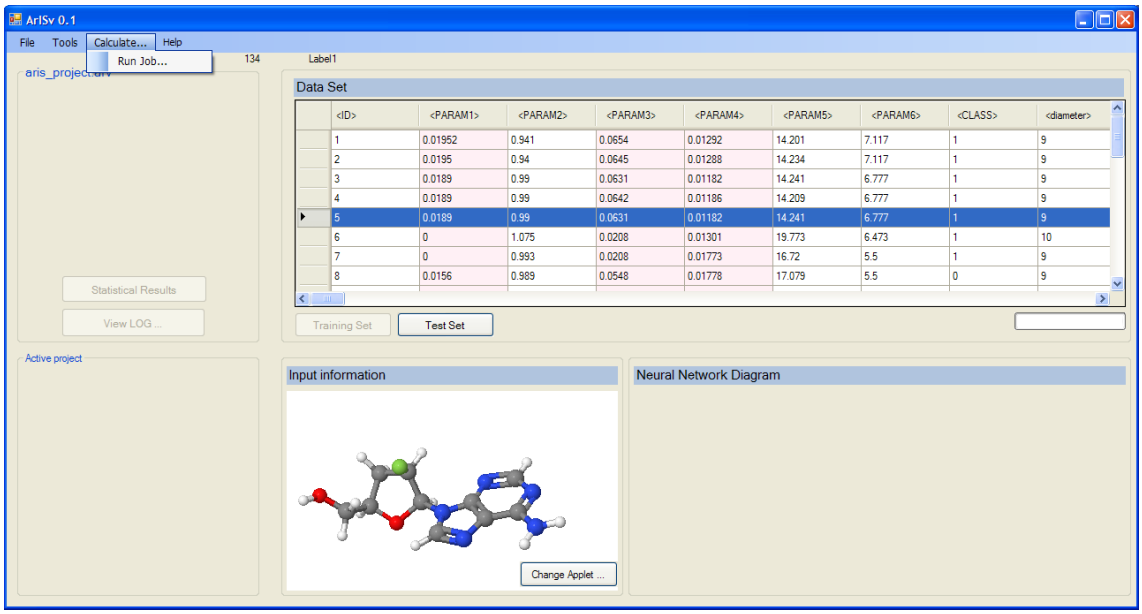


Figura 4.14: *Captura de pantalla de la interfície gràfica d'ArIS: visualització de quimioteques.*

En l'exemple de la figura 4.14 es mostra una quimioteca definida per un conjunt de variables, tres de les quals es troben ombrejades. Si hom clica sobre un descriptor, aquest passa automàticament a utilitzar-se com a senyal d'entrada per al mètode de càlcul, i s'identifica mitjançant aquesta coloració.

Per realitzar un càlcul amb ArISv, és necessari tan sols obrir el panell de càlcul i seleccionar l'opció *Run Job...*. Aleshores es mostra un menú organitzat en etiquetes, que organitzen tota la informació que pot definir-se en un arxiu d'entrada d'ArIS. Per exemple, la figura 4.15 mostra la pestanya corresponent a la definició del mètode de càlcul, on es defineix una xarxa classificatòria de topologia 3-6-4-2 que s'entrenarà amb el mètode BP-*Batch* utilitzant una funció d'activació sigmoïdal.

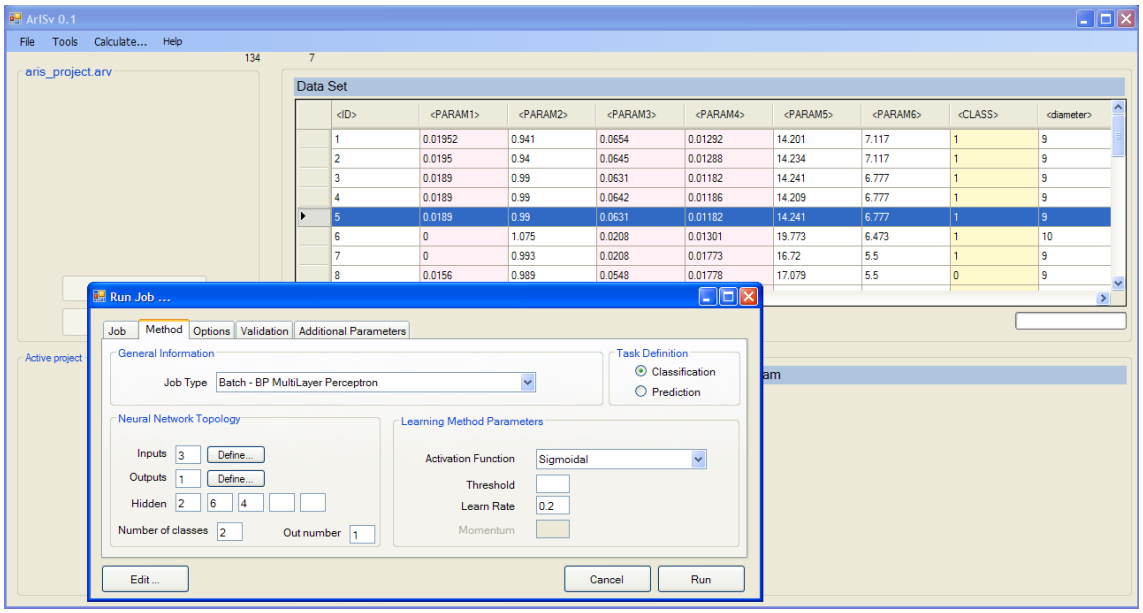


Figura 4.15: *Captura de pantalla de la interfície gràfica d'ArIS: definició d'un càlcul.*

Cal adonar-se que el nombre corresponent a l'etiqueta *Outputs* indica el nombre de descriptors que s'utilitzen per a la definició del valor de sortida (i no el nombre de neurones de la capa de sortida). En aquest exemple en concret és el descriptor `CLASS` el que correspon a la sortida desitjada; es pot observar que aquest descriptor es troba colorat de groc.

Una vegada realitzat el càlcul, els resultats obtinguts s'afegeixen al projecte i apareixen representats com a models en la part esquerra de la finestra principal. En la figura 4.16 es mostra el cas d'un projecte que inclou dos models basats en xarxes neuronals, el primer amb topologia 6-6-2 i el segon amb una 6-4-2.

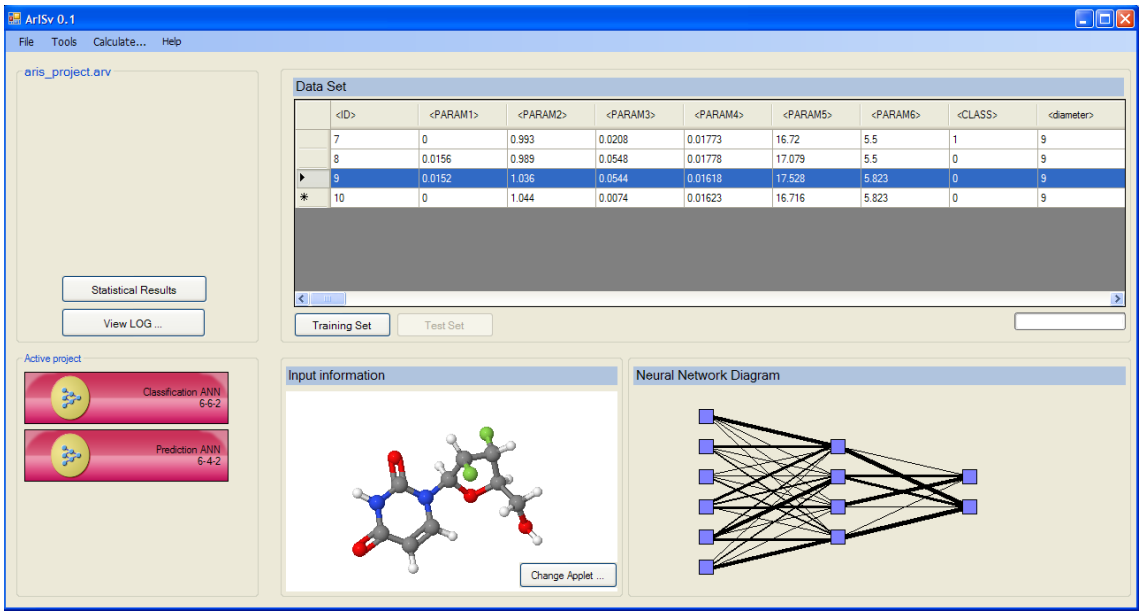


Figura 4.16: *Captura de pantalla de la interfície gràfica d'ArIS: organització dels càlculs realitzats.*

L'anàlisi dels resultats generats es pot realitzar clicant a sobre el model corresponent. En aquest moment es genera un diagrama esquemàtic de la xarxa obtinguda (amb la topologia corresponent, de tal manera que el gruix de les connexions sinàptiques és proporcional al seu pes) i s'habilita el panell amb els resultats de reconeixement i predicció obtinguts, figura 4.17.

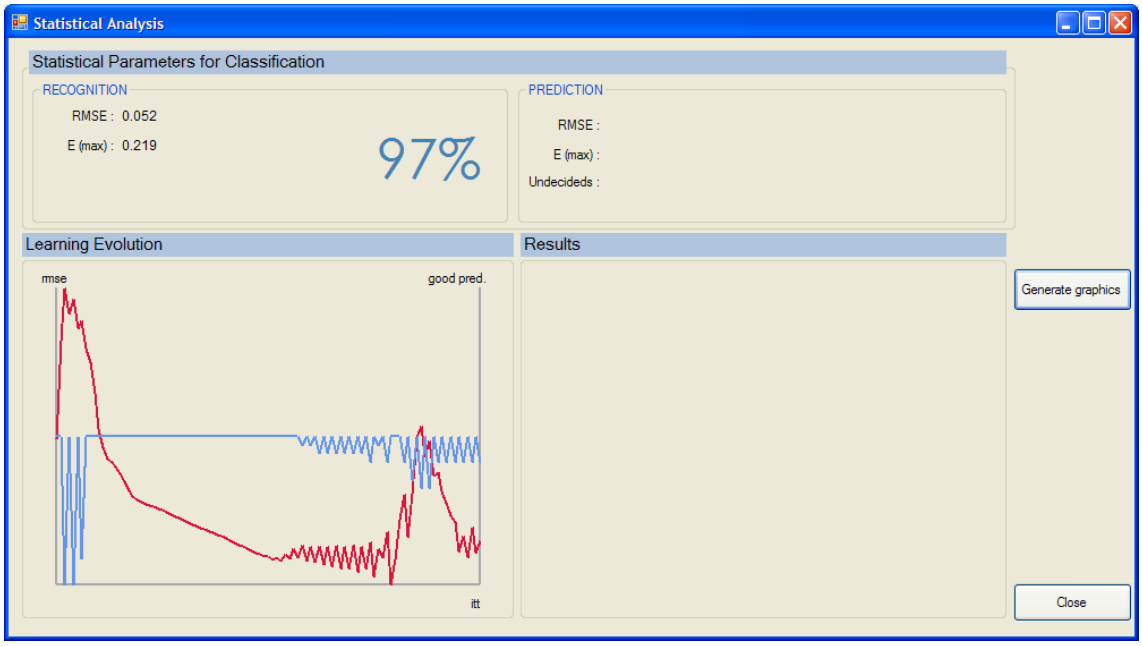


Figura 4.17: *Captura de pantalla de la interfície gràfica d'ArIS: estratègia seguida per mostrar els resultats obtinguts.*

Per tal de millorar la visualització de les dades implicades, es desitja incorporar, en un futur, els SOM i la projecció sobre el pla hiperbòlic per a la representació dels descriptors.

Capítol 5

Mètodes de càlcul disponibles a ArIS

Tal com s'ha comentat en el capítol 3, els algorismes de càlcul que conformen l'eina computacional ArIS estan basats en mètodes d'intel·ligència artificial (IA). Si bé en el capítol 3 s'ha realitzat una breu introducció a alguns d'aquests mètodes, en les següents seccions s'enumeren i es descriuen els pertanyents al programari, parant especial atenció a la seva implementació.

5.1 Perceptrons

Els perceptrons (també anomenats elements de processament) són les unitats bàsiques de les xarxes neuronals. Tanmateix, també poden emprar-se de forma individual (*single perceptron*) per dur a terme tasques de classificació binària de patrons que són linealment separables.

Esquema general

Un perceptró està format per una unitat de processament central que rep la informació a tractar ($\vec{x}$) d'un conjunt de neurones d'entrada $\{x_0, \ldots, x_n\}$, i de la seva combinació n'obté una resposta binària (y_k), indicadora de la classificació, figura 5.1.

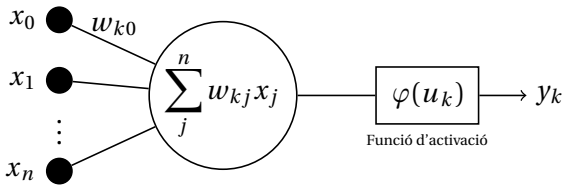


Figura 5.1: *Diagrama esquemàtic d'un perceptró.*

La combinació lineal de $\vec{x}$ es realitza segons el pes que governa cada connexió entre la unitat de processament i la neurona d'entrada corresponent:

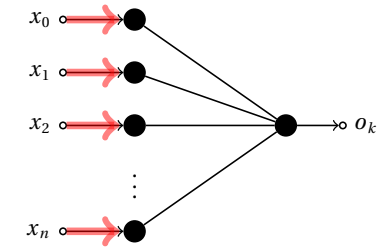
$$u_k = \sum_j^n \cdot w_{kj} x_j \tag{5.1}$$

Per tal que el resultat obtingut de l'equació 5.1 pugui considerar-se indicador d'una classe, cal restringir el seu domini als valors 0 i 1 (cada un d'ells representant una classe). Això s'aconsegueix aplicant una funció d'activació de tipus esglaó (φ) amb un valor llindar (θ) determinat:

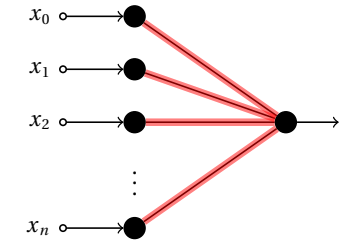
$$\varphi(u_k) = \begin{cases} 1 & \text{si } u_k \geq \theta \\ 0 & \text{si } u_k < \theta \end{cases} \tag{5.2}$$

Algorisme d'aprenentatge

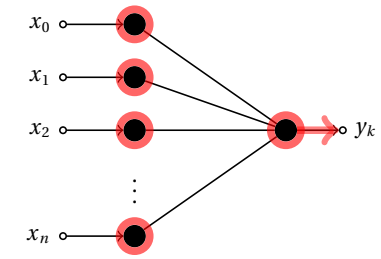
Donat que els pesos sinàptics adequats no són coneguts a priori, cal que siguin optimitzats. En el cas dels perceptrons individuals s'utilitza el mètode d'aprenentatge supervisat anomenat *Perceptron Learning Rule* (PLR) pel seu entrenament. Aquest es basa en l'actualització progressiva dels pesos sinàptics per tal de minimitzar l'error que es comet en la classificació d'un conjunt de vectors entrada (*training set*) amb resultat conegut (o_k), figura 5.2.



La lectura de l'arxiu d'entrada permet obtenir:
1. la topologia de la xarxa (`read_topology`)
2. els valors d'entrenament ($\vec{x}$)
3. les seves respostes respectives, o_k (`read_inputs`)



La definició del tipus de càlcul a realitzar (perceptró) implica la determinació de:
1. la connectivitat entre neurones (`init_weight`)
2. els pesos de cadascuna d'elles (`architecture`)



Una vegada definit tot el sistema, s'inicia el procés d'entrenament (`learning`), en dues etapes:
1. donat $\vec{x}$, es calculen les sortides de cadascuna de les neurones de la xarxa (`Proc_Element`)
2. es compara el resultat obtingut (y_k) amb l'esperat teòricament (o_k)

Figura 5.2: *Representació esquemàtica de l'algorisme de càlcul que segueix un perceptró.*

La modificació que pateixen els pesos en cada iteració és proporcional al producte entre el valor de la neurona d'entrada (x_i) i la diferència entre la resposta predita (y_k) i l'esperada (o_k). El factor de

proporcionalitat rep el nom de *learning rate* ($\varepsilon \in [0,1]$) i determina l'efecte de la correcció sobre el pes actual:

$$w_{ki}(t) = w_{ki}(t-1) + \varepsilon \cdot (o_k - y_k) \cdot x_i(t) \tag{5.3}$$

Amb tot, el mètode d'aprenentatge PLR es pot descriure mitjançant l'algorisme 3.

Algorisme 3 *Perceptron learning rule*, per a classificacions binàries.

Requeriments: $\theta, \mathbb{X} = \{\vec{x_0} \dots \vec{x_m}, \vec{o}\}$

Retorna: $\mathbb{W}$: matriu de pesos optimitzada

1: **for** $x_i \in \mathbb{X}$ **do**
2: **while** convergència = FALSE **do**
3: $u_k \leftarrow \{\vec{x_i}, \mathbb{W}\}$: Càlcul de la combinació lineal
4: $y_k \leftarrow \varphi(u_k)$: Resposta predita
5: **if** $y_k = o_i$ **then**
6: exit
7: **else**
8: $w_{ki}(t)$: Actualització dels pesos
9: **end if**
10: **end while**
11: **end for**

5.1.1 Criteris de convergència

Tal i com s'observa en l'algorisme 3, els mètodes d'optimització de què disposa ArIS requereixen de l'ús d'estructures iteratives. Tot i que normalment hom espera que la resposta millori com més evolucioni, cal disposar de mesures de control que permetin finalitzar el càlcul en el moment òptim. El criteri de convergència és el responsable de determinar si un càlcul ha convergit en una solució al llarg d'un procés iteratiu, i vetlla per què aquesta sigui la millor possible, distingint-la de la resta de possibles solucions.

S'ha incorporat a ArIS un sistema de convergència format per tres condicions:

1. El nombre d'entrades ben predites pel model coincideix amb el nombre de dades aportades en el conjunt d'entrenament.

2. El nombre d'iteracions supera el valor màxim d'iteracions possibles definit per l'usuari.

3. L'error quadràtic mig és menor a un valor llindar definit per l'usuari.

A cada condició se li assigna un codi intern de tal manera que, donat un determinat resultat, l'usuari pot saber quina ha estat la condició que ha conduït a l'aturada del procés iteratiu.

5.1.2 Validació dels perceptrons

La validació del mètode s'ha realitzat amb la classificació d'elements amb dues senyals d'entrada de l'espai $\mathbb{R}$, on els elements amb valors d'entrada negatius pertanyen a la classe 0 i els positius a la classe 1.

Definint un conjunt d'entrenament de tan sols 8 elements (equipoblat en les dues classes), i amb un valor de *threshold* de 0.3, el mètode convergeix després de 10 iteracions i és capaç de predir correctament tot el conjunt d'entrenament. A més, quan s'aplica sobre un conjunt de validació format per 100 elements de cada classe, el mètode els classifica a tots correctament. s

5.2 Mètodes Combinadors Lineals i Adaptatius

5.2.1 ADALINE - ADAptive LINEar neuron

L'ADALINE és un mètode de classificació binària, anàleg als perceptrons, on la resposta s'obté per combinació lineal dels senyals d'entrada. La funció d'activació que s'utilitza en aquest cas és binomial (és a dir, transforma el valor de la combinació lineal en els valors $\{-1,+1\}$) i l'optimització dels pesos durant l'entrenament es realitza abans d'aplicar la funció d'activació, figura 5.3.

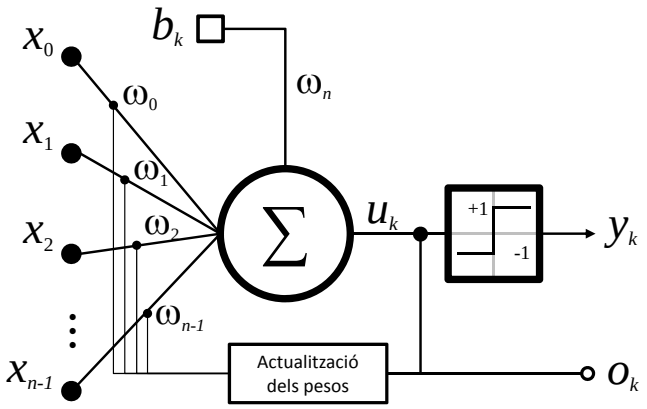


Figura 5.3: *Representació esquemàtica del mètode de classificació ADALINE.*

El mètode *Least-Mean-Square*

El mètode *Least-Mean-Square* (LMS) té com a objectiu trobar la combinació de pesos que minimitza la funció d'error corresponent a la definició de l'error quadràtic mitjà comès en les prediccions realitzades sobre el conjunt de tots els patrons d'entrenament:

$$\mathscr{E} = \frac{1}{L}\sum_{k=1}^{L}(o_k - u_k)^2 \tag{5.4}$$

L'expressió de la funció d'error es pot desenvolupar en forma matricial:

$$\begin{aligned}\mathscr{E} &= \frac{1}{L}(\mathbb{O}-\mathbb{X}\mathbb{W})^T(\mathbb{O}-\mathbb{X}\mathbb{W}) \\ &= \frac{1}{L}\left(\mathbb{O}^T-(\mathbb{X}\mathbb{W})^T\right)(\mathbb{O}-\mathbb{X}\mathbb{W}) \\ &= \frac{1}{L}\left[\mathbb{O}^T\mathbb{O}-\mathbb{W}^T\mathbb{X}^T\mathbb{O}-\mathbb{O}^T\mathbb{X}\mathbb{W}+\mathbb{W}^T\mathbb{X}^T\mathbb{X}\mathbb{W}\right] \end{aligned} \tag{5.5}$$

Per tal de localitzar el mínim de la funció d'error hom pot emprar informació del gradient de la funció en cada punt. Per la condició de mínim, la derivada de l'error respecte dels pesos ha de ser zero en aquest punt:

$$\frac{\partial \mathscr{E}}{\partial \mathbb{W}} = \bigtriangledown \mathscr{E} = \frac{1}{L}\left[-\mathbb{X}^T\mathbb{O}-\mathbb{O}^T\mathbb{X}+2\mathbb{X}^T\mathbb{X}\mathbb{W}\right] = 0 \tag{5.6}$$

d'aquesta expressió es pot deduir que el gradient esdevé zero quan es compleix la igualtat:

$$\mathbb{X}^T\mathbb{X}\mathbb{W} = \mathbb{X}^T\mathbb{O} \tag{5.7}$$

A més, el sistema d'equacions que estableix l'eq. 5.7 és un sistema compatible i determinat, ja que correspon a l'ajust per mínims quadrats del sistema d'equacions derivat de mostrar el conjunt d'entrenament al sistema neuronal:

$$\mathbb{X}\mathbb{W} = \mathbb{O} \tag{5.8}$$

El càlcul del vector de pesos solució es pot realitzar, aleshores, mitjançant un dels diferents mètodes de resolució de sistemes d'equacions lineals.

Widrow-Hoff delta rule

De forma alternativa, la determinació del millor conjunt de pesos que minimitzen l'eq. 5.4 es pot realitzar utilitzant mètodes d'optimització. Disposant d'informació referent al gradient de la funció en cada punt, el mètode d'entrenament anomenat *Widrow-Hoff delta rule* (o regla δ) segueix una estratègia *steepest descent* per millorar de forma iterativa el vector de pesos:

$$w_{ij}(t) = w_{ij}(t-1) - \Delta w_{ij} \tag{5.9}$$

on,

$$\Delta w_{ij} = -\eta \frac{\partial \mathscr{E}}{\partial w_{ij}} \tag{5.10}$$

que, combinat amb el resultat de l'eq. 5.6 per a un pes w_{ij} determinat, resulta:

$$w_{ij}(t) = w_{ij}(t-1) + 2\eta x_i(o_k - y_k) = w_{ij}(t-1) + 2\eta x_i \delta \tag{5.11}$$

La correcció es realitza de forma seqüencial per tots els exemples mostrats a la xarxa, de manera

que per un vector d'entrada donat, cadascun dels pesos són optimitzats abans de passar al següent vector.

Comparació entre mètodes unineuronals

Tant ADALINE com el model perceptró serveixen per a la classificació de patrons separables linealment. Per aquest motiu es compara la utilització d'ambdós mètodes per un mateix problema: la classificació binària de vectors bidimensionals.

Es crea un llistat de 50 punts en el pla $\mathbb{R}^2$, de posicions (x, y) aleatòries, els quals es classifiquen segons la recta $y = 5 - 3x$, definida de forma arbitraria. Les dades es presenten a un perceptró i a un ADALINE, els quals s'entrenen corresponentment.

Mitjançant el perceptró el 98% dels punts es classifiquen correctament, mentre que amb l'ús d'ADALINE se'n classifiquen el 84%, figura 5.4.

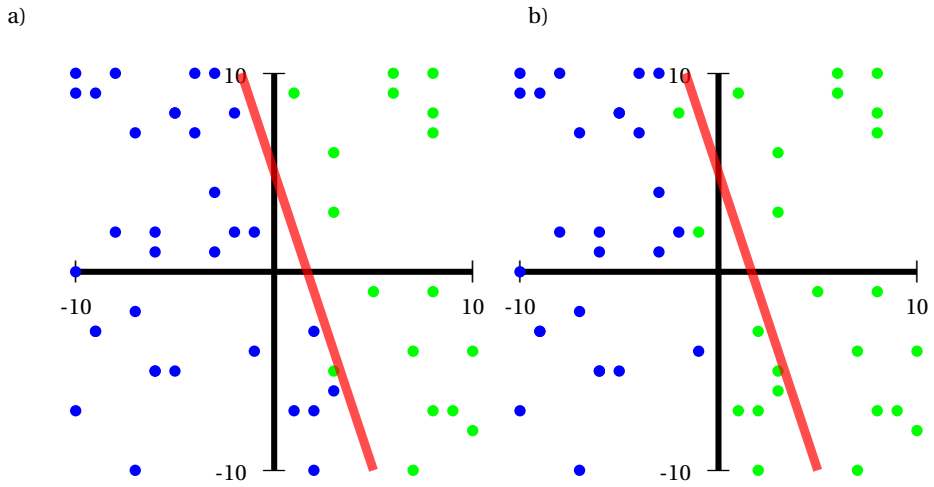


Figura 5.4: *Resultats obtinguts en la classificació segons la recta $y = 5 - 3x$ de 50 punts generats aleatòriament utilitzant perceptrons (a) o ADALINE (b).*

5.2.2 MADALINE - Multiple ADAptive LINEar neuron

Havent demostrat que ADALINE és capaç de distingir patrons linealment separables, el consegüent pas és avaluar el comportament en interconnectar diferents unitats de processament. Aquestes tindrien la capacitat de separar patrons per més d'un hiperpla?

La incorporació de neurones internes

Les estructures formades per més d'una ADALINE són anomenades múltiples ADALINE o MADALINE. Conformen la primera aproximació a les xarxes neuronals artificials de més d'una capa de neurones.

La gran innovació d'aquest model és la incorporació d'un conjunt de neurones entre la capa de neurones d'entrada i les neurones de sortida, que poden estar alhora agrupades en forma de capes

(*hidden layers*). L'augment del volum d'informació intercanviada entre neurones permet millorar la capacitat de separació de patrons allunyant-la del que era estrictament els patrons lineals.[78]

Donat que cada node present retorna una sortida corresponent al resultat d'una ADALINE (és a dir un valor de {-1, +1}), aquest mètode es començà a utilitzar pel reconeixement visual,[79] on cada píxel correspon a una unitat ADALINE (figura 5.5) . Aquesta estructura permet ampliar a més de dues classes l'objectiu de les classificacions, havent-hi tantes neurones de sortida com classes es desitgi classificar.

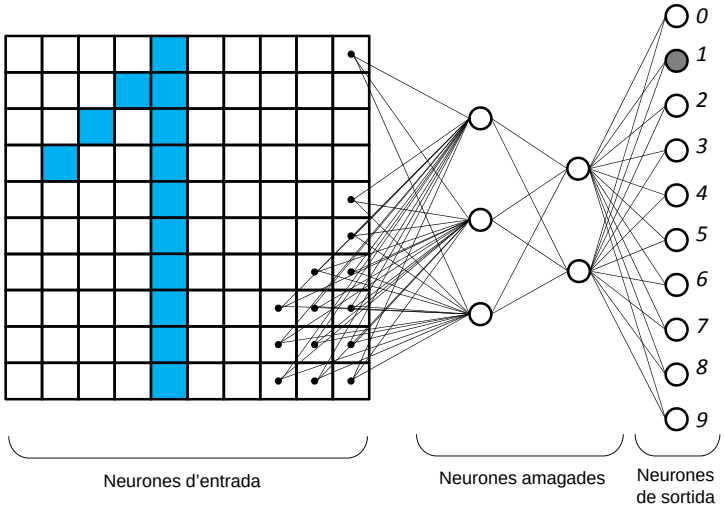


Figura 5.5: *Representació esquemàtica de l'aplicació del reconeixement de dígits mitjançant MADALINE. La capa de neurones d'entrada està formada per un panell de 10x10 píxels, cadascun dels quals correspon a una neurona. Tots ells es troben connectats amb totes les neurones de la primera capa amagada, que a la vegada ho estan amb les neurones de la segona capa amagada. El resultat de les neurones amagades s'acaben combinant en la capa de sortida, que en aquest cas està formada per 10 neurones, on cada una d'elles correspon a un dígit diferent.*

Un mètode d'entrenament propi

La presència de capes amagades implica un canvi fonamental en el mètode d'entrenament, doncs ha d'incorporar els pesos corresponents a les connexions entre aquestes neurones. Això fa que no es pugui aplicar l'algorisme LMS (pàg. 104) per dos motius:

1. Encara que es podria aplicar la regla δ per a les neurones de sortida, ja que es disposa del resultat esperat, aquest no es pot aplicar en les neurones amagades perquè no es té coneixement del valor esperat per elles.

2. El mètode LMS actua sobre el resultat del combinador lineal, no sobre les sortides bipolars d'ADALINE

Per aquests motius, MADALINE utilitza un mètode propi d'entrenament seqüencial anomenat MRII[80] (algorisme 4). Que sigui seqüencial vol dir que l'entrenament es realitza sobre un input en concret, i la matriu de pesos obtinguda de l'entrenament d'un input es considera la matriu de pesos inicials pel següent.

Algorisme 4 Mètode MRII d'entrenament per a xarxes del tipus MADALINE.

Requeriments: $\mathbb{W}$:matriu de pesos inicials aleatoris

while input ∈ {EPOCH} **do**
  $\mathscr{E}(\mathbb{W}) \leftarrow$ `bad_outputs`: avaluació del nombre de mal classificats
  while layer ∈ {input, hidden, output} **do**
    while i ∈ {layer} **do**
      ACCEPTED ← 0
      repeat
        M ← min$\{u_i\}$: se selecciona la neurona M amb menor valor de sortida
        repeat
          $\mathbb{W}' \leftarrow$ `change_weights`: actualització dels pesos
        until flip(y(M))
        $\mathscr{E}(\mathbb{W}') \leftarrow$ `bad_outputs`
        if $(\mathscr{E}(\mathbb{W}') > \mathscr{E}(\mathbb{W}))$ **o** $(\mathscr{E}(\mathbb{W}') = 0)$ **then**
          $\mathbb{W}' = \mathbb{W}$
          ACCEPTED←1
        end if
      until ACCEPTED=0
    end while
    if $\mathscr{E}(\mathbb{W}) = 0$ **then**
      next i
    end if
  end while
end while
return $\mathbb{W}$

L'actualització dels pesos es realitza de tal manera que es mantingui el principi de mínima pertorbació (*minimal disturbance*). Tenint en compte la funció d'activació binomial (φ), centrada en zero, aplicada després de la combinació lineal, s'obté:

$$y_k = \varphi(u_k) = \text{sign}(u_k) = \text{sign}\left(\sum w_{kj} \cdot x_j\right) \tag{5.12}$$

De manera que, si hom desitja intercanviar la sortida d'una neurona M, que és originàriament negativa, caldrà augmentar el valor de la combinació lineal. Això es realitza de forma anàloga al mètode LMS en ADALINE, mitjançant la regla δ:

$$w(t) = w(t-1) + \eta \cdot x_i \cdot \delta \tag{5.13}$$

on δ correspon a la diferència entre el valor desitjat (o_M) i l'obtingut (y_M).

En el cas d'haver canviat totes les possibles neurones de la xarxa amb $y \rightarrow 0$ i no haver assolit un error nul ($\mathscr{E}$, corresponent al nombre d'entrades mal classificades), es pot aplicar l'aproximació *Pairwise Trial Adaptations*. Aquesta consisteix en tornar a repetir el procés MRII d'entrenament, però

aplicant la correcció dels pesos sobre les 2 neurones amb valors de sortida menors en comptes d'un. Fins i tot, si aquest continua essent insuficient, es pot ampliar a un nombre major de neurones de forma progressiva.[80]

Validació de MADALINE

Donat que aquest mètode fou utilitzat (en els seus inicis) pel reconeixement visual de patrons, es proposa el reconeixement dels dígits {0,1,2} per a la seva validació. L'estructura de la xarxa és anàloga a la descrita a la figura 5.5, on la capa de neurones d'entrada està formada per un centenar la píxels disposats de forma 10x10. Es defineixen dues capes amagades amb 30 i 10 neurones cadascuna d'elles. La capa de sortida disposa de tres neurones (una per a cada classe), de manera que s'espera que als patrons amb classe 0 els correspongui el vector de sortida $(+1,-1,-1)$, $(-1,+1,-1)$ als de classe 1 i $(-1,-1,+1)$ a la classe 2 (cal recordar que la resposta de cada unitat de processament és en aquest cas binomial, figura 5.5 pàg. 107).

Es crea una base de dades d'entrenament amb 15 dígits escrits a mà sobre la malla quadriculada (figura 5.6), amb 5 exemples de cada dígit. Els píxels que queden ombrejats reben un valor d'entrada unitari, mentre que a la resta se'ls assigna el valor zero.

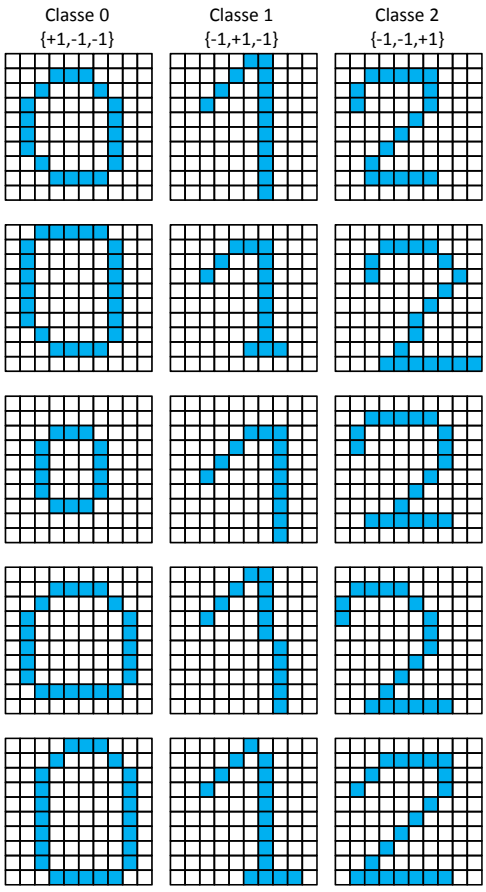


Figura 5.6: *Conjunt d'exemples utilitzats pel reconeixement de dígits mitjançant MADALINE.*

D'aquesta manera, cada vector d'entrada està format per 100 posicions que poden prendre els valors {0,+1}. L'entrenament mitjançant MADALINE permet reconèixer correctament el 100% dels 15 exemples mostrats a la xarxa, i de tots els exemples considerats posteriorment com a conjunt de validació.

Identificació de derivats tetrapirròlics

El funcionament d'aquest tipus de xarxa s'ha validat amb la classificació de fotosensibilitzadors tetrapirròlics, segons característiques estructurals. Es crea una quimioteca de 14 derivats de porficè, 58 derivats de porfirina, 32 ftalocianines i 21 clorines. Tots ells constitueixen fotosensibilitzadors d'interès per a la teràpia fotodinàmica (part III, pàg. 145), i es caracteritzen per petits canvis estructurals en el seu nucli central, figura 5.7. Donada la seva importància, i aprofitant les cerques bibliogràfiques realitzades, es planteja la seva classificació segons la seva estructura.

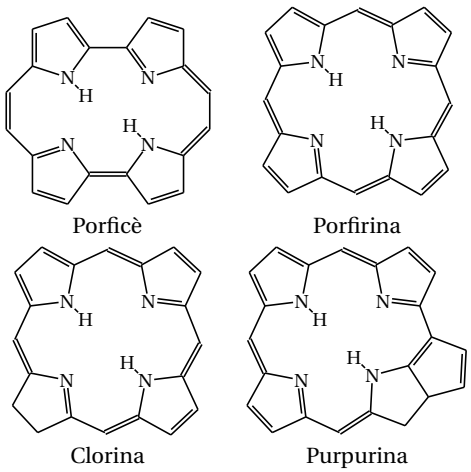


Figura 5.7: *Estructura molecular dels quatre nuclis dels derivats tetrapirròlics considerats en la validació de MADALINE.*

La descripció de l'estructura dels FS es realitza mitjançant els *fingerprints* (secció 2.1.1, pàg. 53) de tipus BITMACCS disponibles en el programari MOE2009.10.[81] Aquests permeten codificar les característiques estructurals dels derivats tetrapirròlics en un vector binari (format per 165 posicions) que és utilitzat com a vector d'entrada en la xarxa neuronal artificial (el vector de sortida conté quatre posicions, una per cada classe considerada).

Utilitzant les estructures moleculars dels fotosensibilitzadors definits en la part III (pàg.145) del treball, es defineix un conjunt d'entrenament amb 105 FS, i 72 (aproximadament el 40% del total, seleccionats de forma aleatòria) per a la validació externa. La distribució dels nuclis en cadascun dels conjunts es mostra a la taula 5.1.

Taula 5.1: *Distribució de la base de dades de FS segons el seu nucli.*

	Porfirines	**Porficens**	**Clorines**	**Purpurines**	**TOTAL**
Entrenament	15	60	15	15	105
Validació	8	49	8	7	72

El model obtingut amb MADALINE, per a la classificació dels quatre nuclis de derivats tetrapirròlics, permet reconèixer correctament el 80% dels FS definits en l'entrenament i presenta una capacitat de predicció del 65% per al conjunt de validació externa.

Per tal d'incorporar en aquesta discussió el mètode ADALINE, s'avalua la reducció del problema a la identificació de derivats porficènics (60 FS per l'entrenament i 49 per a la validació) de la resta de FS (45 per l'entrenament i 23 per a la validació), ja que ADALINE només és capaç de donar com a resultat dos valors de sortida. Per aquesta classificació, ADALINE divergeix mentre que MADALINE assoleix un reconeixement del 93% i una predicció del 82%, fet que demostra de la potència de poder presentar capes amagades de neurones.

D'altra banda, la comparació dels resultats amb una i dues capes amagades per a MADALINE permet determinar l'efecte de la topologia: com anteriorment, la classificació arriba al 93% de reconeixement i al 82% de predicció amb dues capes amagades i es queda en un reconeixement més moderat (88%, amb una capacitat de predicció similar) amb una única capa amagada.

5.3 Xarxes neuronals basades en perceptrons i amb més d'una capa

5.3.1 L'algorisme de retropropagació de l'error

Aquest algorisme, també anomenat *Back-propagation learning rule* (BP), és el mètode d'entrenament supervisat emprat normalment per a les xarxes multicapa (BP-ANN), on la funció d'activació és una funció sigmoïdal. Utilitza essencialment una generalització de la regla δ (pàg. 105), distingint entre les neurones de la capa de sortida d'aquelles que conformen les capes amagades. S'aplica un mètode d'optimització basat en *steepest descent*, estimant el valor del gradient en cada punt.

La funció d'error a minimitzar es defineix com l'error quadràtic mig sobre el conjunt de totes les neurones de sortida (C):

$$\mathscr{E} = \frac{1}{2}\sum_{i\in C}\left(o_i - y_i\right)^2 = \frac{1}{2}\sum_{i\in C} e_i^2 \tag{5.14}$$

Càlcul de la resposta de les neurones de sortida

Correspon al cas anàleg a l'entrenament LMS. En conèixer el valor de sortida esperat per a cadascuna de les neurones, es pot estimar l'error comès com la diferència dels dos valors, podent aplicar l'eq. 5.14. El factor corrector dels pesos sinàptics es realitza després d'aplicar la funció d'activació, de manera que cal tenir en compte la transformació no lineal a l'hora d'estudiar com varia l'error en funció d'un pes. El gradient local per a un pes sinàptic donat (w_{ji}) esdevé, aplicant la regla de la cadena:

$$\frac{\partial \mathscr{E}}{\partial w_{ji}} = \frac{\partial \mathscr{E}}{\partial e_j}\cdot\frac{\partial e_j}{\partial y_j}\cdot\frac{\partial y_j}{\partial u_j}\cdot\frac{\partial u_j}{\partial w_{ji}} \tag{5.15}$$

Tenint en compte la definició de l'error anterior (eq. 5.14) i utilitzant les equacions 3.1 (pàg. 63) i 3.4 (pàg. 64) per al càlcul dels resultats de sortida de cadascuna de les neurones, s'obté:

$$\begin{aligned} &= \frac{\partial}{\partial e_j}\left(\frac{1}{2}\sum e_j^2\right)\cdot\frac{\partial}{\partial y_j}\left(o_j - y_j\right)\cdot\frac{\partial \varphi}{\partial u_j}\cdot\frac{\partial}{\partial w_{ji}}\left(\sum w_{jk}y_k\right) \\ &= \left(e_j\right)\cdot(-1)\cdot\left(\varphi'\right)\cdot\left(y_i\right) \end{aligned} \tag{5.16}$$

Així doncs, la correcció del pes sinàptic (Δw_{ji}) definit per la regla δ correspon al producte anterior pel *learning rate*, η:

$$\Delta w_{ji} = -\eta\cdot\left(-e_j\cdot\varphi_j'\right)\cdot y_i \tag{5.17}$$

Normalment s'agrupen els termes corresponents a l'error comès i a la derivada de la funció d'activació en forma d'un sol paràmetre (δ_j), per tal d'expressar el terme corrector com un producte entre el què correspondria al gradient local i al senyal d'entrada:

$$\Delta w_{ji} = \eta\cdot\delta_j\cdot y_i \tag{5.18}$$

Càlcul de la resposta de les neurones amagades

A diferència del cas anterior, en aquest no es disposa del valor esperat. El valor de δ_j, per una neurona j qualsevol de les capes amagades o de sortida, es defineix com:

$$\delta_j = -\frac{\partial \mathcal{E}}{\partial y_j}\cdot\frac{\partial y_j}{\partial u_j} \tag{5.19}$$

Aplicant la definició de l'error (eq. 5.14), de forma anàloga al raonament seguit en el desenvolupament eq. 5.16, hom pot demostrar que la variació de l'error en funció del valor de sortida y_j queda definit segons l'eq. 5.20 i és funció de les δ de les neurones (δ_k) que reben el resultat de la neurona j com a entrada.

$$\frac{\partial \mathcal{E}}{\partial y_j} = -\sum_k \delta_k w_{kj} \tag{5.20}$$

Així doncs, el mètode de retropropagació permet mantenir la definició del terme corrector tant per les neurones de sortida com per a les amagades (eq. 5.18), on l'únic que canvia és la manera com es calcula el valor de δ (taula 5.2).

Taula 5.2: *Generalització de la regla δ; diferència entre el càlcul del terme corrector en neurones amagades i de sortida.*

$$\Delta w_{ji} = \eta\cdot\delta_j\cdot y_i$$

Neurones amagades	Neurones de sortida
$\delta_j = \varphi_j'\cdot\sum_k \delta_k\cdot w_{kj}$	$\delta_j = e_j\cdot\varphi_j'$

La funció d'activació

La funció d'activació utilitzada en l'entrenament de retropropagació ha de ser una funció contínua (eq.5.16). Atenent a aquest fet, el programari ArIS disposa de dues possibles funcions d'activació per l'entrenament supervisat de xarxes multicapa: una funció sigmoïdal i la tangent hiperbòlica.

El tipus de funció d'activació es pot definir dins l'arxiu d'entrada amb el comandament `#Active Function`. Si bé mètodes com ADALINE i MADALINE només accepten un determinat tipus de funció d'activació, la taula 5.3 recull tots els valors que pot prendre la paraula clau anterior, reservant el valor zero pels casos que no es desitja aplicar cap tipus de funció d'activació. Més que el requeriment d'un mètode en concret, el valor nul constitueix una estratègia per millorar la generalització del codi implementat. La descripció de cadascuna de les funcions d'activació disponibles es troba dins l'apartat 3.1.2, pàg. 63.

Taula 5.3: *Llistat del tipus de funcions d'activació disponibles en ArIs.*

ID.	Funció	Mètodes disponibles
1	Lineal	Perceptrons, BP-ANN, GNN
2	Binària	Perceptrons
3	Sigmoidal	BP-ANN, GNN
4	Bipolar	ADALINE, MADALINE
5	Tangent hiperbòlica	BP-ANN, GNN

Modes d'entrenament

El mètode BP d'entrenament ha de permetre adaptar les connexions sinàptiques per tots els vectors implicats. Això significa que tots i cadascun d'ells han de passar en algun moment per l'ANN. Aquest procés es pot realitzar de dues maneres diferents:

1. Presentant els vectors d'entrada un rere l'altre, aplicant el mètode BP per a cadascun d'ells (anomenat mode seqüencial, figura 5.8):

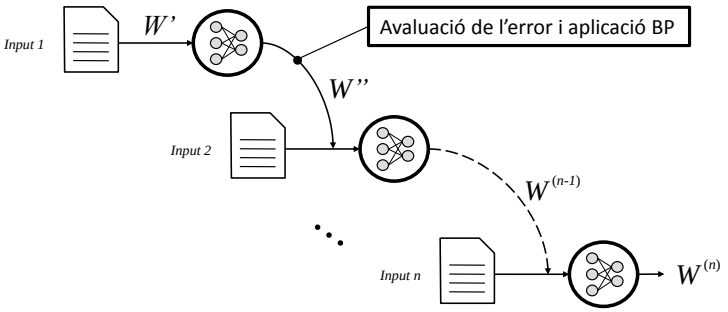


Figura 5.8: *Il·lustració del mode seqüencial d'entrenament de l'algorisme BP. Es realitza un entrenament individual per cada vector d'entrada, on la matriu de pesos resultant d'un exemple és utilitzada com a punt inicial pel següent.*

2. Avaluant l'error comès en el conjunt de tots els vectors que formen el conjunt d'entrenament

(*epoch*), per a una matriu de pesos donada, i aprofitar aquest resultat per actualitzar els pesos segons l'algorisme BP (que rep el nom de mode en lots o *batch*, figura 5.9):

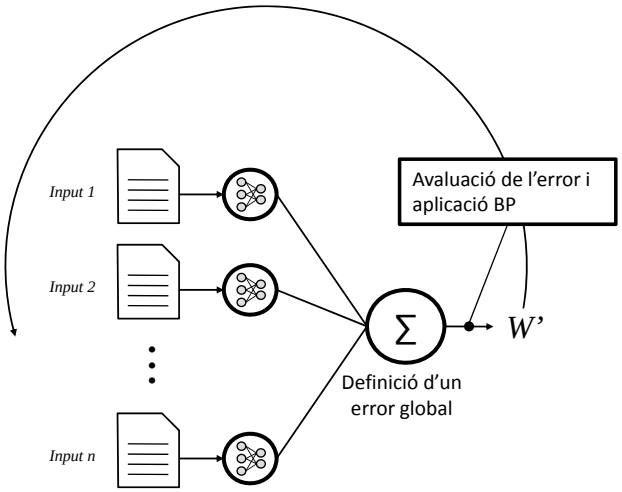


Figura 5.9: *Il·lustració del mode d'entrenament en lots de l'algorisme BP. Tots els vectors d'entrada són avaluats per una mateixa matriu de pesos, la qual s'actualitza a posteriori i dins d'un procés iteratiu fins assolir algun dels criteris de convergència establerts.*

Algorisme 5 Mètode d'aprenentatge supervisat *Back-Propagation* Seqüencial.

Requeriments: $\mathbb{W}$:matriu de pesos inicials aleatoris

while CONVERGENCE=FALSE **do**
  while input $\in$ {EPOCH} **do**
    while neuron $\in$ {output layer} **do**
      $e^{\text{input}}_{\text{neuron}} \leftarrow o_{\text{neuron}} - y_{\text{neuron}}$
      $\delta_t \leftarrow -e_t \cdot \varphi'_t$
    end while
    while s $\in$ {hidden layer} **do**
      $\delta_s \leftarrow -\varphi'_s \cdot \sum_k \delta_k \cdot w_{ks}$
    end while
  end while
  Actualització dels pesos: $w_{ji} \leftarrow w_{ji} - \eta \delta_j y_i$
end while

La correcció corresponent al mode seqüencial es realitza segons l'equació 5.18 (algorisme 5). Tanmateix, la implementació del mode *batch* requereix d'una lleugera modificació del codi, per tal d'encabir-hi la nova definició de la funció d'error quadràtic mig ($\mathscr{E}_{av}$) aplicada a un conjunt de vectors d'entrenament (*epoch*) format per N exemples i C neurones de sortida:

$$\mathscr{E}_{av} = \frac{1}{2N} \sum_{i \in N} \sum_{j \in C} e_j^2 \tag{5.21}$$

Encara que la redefinició de l'error implica alhora una nova definició del paràmetre δ, l'expressió de la correcció dels pesos sinàptics (Δw, eq. 5.18) pot continuar essent vàlida si s'adapta el codi convenientment i així incloure el mode *batch* en el càlcul de δ (algorisme 6). Així doncs, en el procés d'*steepest descent*, l'actualització d'un pes, després de passar les N entrades de l'*epoch*, es realitza segons l'expressió:

$$w_{ji}(t) = w_{ji}(t-1) + \frac{1}{N}\left(\eta \sum_{k}^{N} \delta_j^k y_i^k\right) \tag{5.22}$$

Algorisme 6 Mètode d'aprenentatge supervisat *Back-Propagation Batch*.

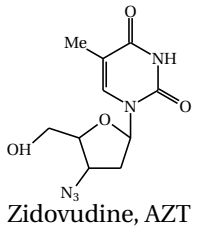
Requeriments: $\mathbb{W}$:matriu de pesos inicials aleatoris

while CONVERGENCE=FALSE **do**
  rtat() $\leftarrow$ `test_set`: càlcul del resultat per a tots els vectors d'entrenament
  $\mathscr{E}_{av} = 0$
  while input $\in$ {EPOCH} **do**
    while neuron $\in$ {output layer} **do**
      $e_{neuron}^{input} \leftarrow o_{neuron} - y_{neuron}$
      $\mathscr{E}_{av} \leftarrow \mathscr{E}_{av} + \left(e_{input}^{neuron}\right)^2$
    end while
  end while
  $\mathscr{E}_{av} = \mathscr{E}_{av}/2N$
  while t $\in$ {output layer} **do**
    $\delta_t \leftarrow -e_t \cdot \varphi_t'$
  end while
  while s $\in$ {hidden layer} **do**
    $\delta_s \leftarrow -\varphi_s' \cdot \sum_k \delta_k \cdot w_{ks}$
  end while
  $w_{ji} \leftarrow w_{ji} - \eta \delta_j y_i$
end while

5.3.2 Predicció de l'activitat anti-VIH d'inhibidors de la transcriptasa inversa

Es realitza prenent com a referència un estudi de Tetko *et al.* on s'avalua la capacitat de classificar un conjunt d'inhibidors de la transcriptasa inversa (TI) del virus de la immunodeficiència humana (VIH) amb ANN, mitjançant l'algorisme d'entrenament BP-Batch.[82] La quimioteca es descriu mitjançant 6 descriptors topològics, corresponents a les modificacions de 3 índexos de Kier (que descriuen aspectes de la forma molecular), l'índex de Balaban i el nombre de Wiener (com estimació de la mida molecular) i un descriptor estructural[83] (que permet la comparació de sistemes moleculars).

Per tal d'obtenir uns resultats comparables amb l'estudi bibliogràfic que s'ha pres de referència, es considera el mateix conjunt de 44 inhibidors de la TI anàlegs a l'AZT (**1**), i es divideixen de la mateixa manera: un conjunt d'entrenament (format pels mateixos 30 compostos) i un conjunt de validació externa (amb 14 compostos). A més, es respecten els mateixos sis descriptors utilitzats en l'establiment del model de predicció bibliogràfic. El millor model trobat pels autors del sistema definit pels sis descriptors correspon a una xarxa de topologia 6-10-2, que presenta un reconeixement del 100% (30/30) i una capacitat de predicció del 71% (10/14).



Zidovudine, AZT

1

Amb la topologia proposada bibliogràficament, l'aplicació d'ArIS permet igualar tant la taxa de reconeixement com la de predicció dels resultats descrits. Per això és necessari definir un *learning rate* de 0.1 i un *momentum rate* de 0.5, i deixar evolucionar l'algorisme BP fins a un màxim de 10000 iteracions.

Escalat dels valors d'entrada

L'exemple proposat per a la validació del BP és el primer en el qual la naturalesa dels descriptors emprats és heterogènia, i el marge de valors que poden prendre uns i altres és molt diferent. Per aquest motiu, hom troba la necessitat d'incorporar mètodes d'escalat de les dades per tal d'aconseguir assolir els resultats bibliogràfics.

Sense la incorporació d'aquest element els resultats que es poden aconseguir queden molt lluny dels anteriors, i en el millor dels casos tan sols s'aconsegueix un reconeixement del 70% i una taxa de predicció del 64%, mantenint les mateixes condicions estructurals i d'entrenament que en el cas anterior.

Aquests resultats posen en evidència la necessitat d'incorporar mètodes per a l'escalat dels vectors d'entrada. S'implementen els mètodes d'escalat lineal i per interval dins del programari ArIS(apartat 4.2, pàg. 85). A partir d'aquest moment, en tots els models que es deriven de l'aplicació de l'algorisme de retropropagació de l'error, s'ha considerat l'escalat de les dades d'entrada com a pas previ.

Mètodes de validació interna

Si bé l'aplicació del mètode BP-Batch permet obtenir resultats comparables als bibliogràfics, aquests no s'han pogut obtenir amb el mode seqüencial. Tot i quedar-se molt a prop del millor resultat bibliogràfic, el mode BP-Seqüencial presenta un reconeixement del 100% però una predicció del 64% (9/14), que no s'ha pogut superar amb la modificació de paràmetres estructurals.

No sorprèn que el mètode com es presenten els vectors d'entrenament a la xarxa esdevingui fonamental per a l'obtenció de bons resultats, ja que dependrà de cada aplicació en concret que un

mode d'entrenament sigui més o menys adequat.[48] Tot i així, els resultats obtinguts condueixen a la possibilitat d'incorporar altres estratègies de presentació dels vectors d'entrenament, que ajudin a millorar l'adaptació dels pesos sinàptics al conjunt de tots els vectors.

Aquest és el moment en què s'implementen dins ArIS els diferents mètodes de validació interna (apartat 4.3, pàg. 85), els quals intenten millorar la capacitat de predicció del model separant un cert nombre d'elements del conjunt d'entrenament per simular el seu comportament sobre patrons desconeguts.

Amb la incorporació de l'*split-half validation* s'aconsegueix millorar els resultats bibliogràfics, obtenint un reconeixement del 100% i una predicció de prop del 86% (12/14).

5.3.3 Predicció de la permeabilitat de la barrera hematoencefàlica

El sistema nerviós central es troba físicament separat dels vasos sanguinis com a mètode de protecció front a substàncies que li poden ser tòxiques. Els vasos capil·lars que accedeixen al cervell tenen associat un endoteli amb un major contingut de lípids en la membrana, fet que restringeix el pas de molècules hidrosolubles, que es coneix amb el nom de barrera hematoencefàlica (*blood-brain barrier*, BBB).

S'ha estudiat la capacitat de l'algorisme BP a entrenar ANN amb capacitat de predir si un determinat fàrmac és capaç de travessar la barrera hematoencefàlica o no, i comparar-lo amb estudis computacionals ja existents.[84] Per a l'entrenament i validació de l'ANN s'utilitza un conjunt de 80 molècules de la base de dades Maybridge[85] de les quals es disposa la informació desitjada. Aquestes se seleccionen de forma aleatòria, fent que el 45% de les molècules seleccionades no travessen la BBB.

Es calculen 184 descriptors que no tenen en compte la informació conformacional de les molècules mitjançant el programari MOE2007.09,[86] sobre els quals es realitza una anàlisi de components principals. El 75% de la variància de les dades és explicada per les 7 primeres components principals, que s'utilitzen com a valors d'entrada de la xarxa.

La quimioteca es divideix en en un conjunt d'entrenament, format per 72 compostos (12 dels quals s'utilitzen per a la validació interna), i un conjunt de validació externa format pels 8 compostos restants.

Seguint un procés de prova i ajust, s'estudien tant els paràmetres estructurals de la xarxa (*learn rate*, nombre de capes i neurones amagades, iteracions màximes, etc.) com el nombre de neurones d'entrada (s'avalua l'efecte de la supressió successiva de les components principals amb menor variància explicada). Una vegada identificat el millor model (establert mitjançant BP-Batch), es compara amb el seu anàleg entrenat amb l'algorisme BP-Seqüencial, taula 5.4.

Taula 5.4: *Resultats obtingut per a la predicció de la permeabilitat BBB.*

Mètode	Topologia	η	% Entrenament	Val. interna	Val. externa
BP-Batch	7-10-5-2	0.1	100.0	87.5	
	6-10-4-2	0.1	100.0	100.0	100.0
	5-10-4-2	0.1	100.0	100.0	100.0
BP-Seqüencial	8-10-4-2	0.1	100.0	75.0	75.0

Els resultats obtinguts en qualsevol de les topologies mostrades en la taula 5.4 es troba considerablement per sobre dels resultats bibliogràfics presos de referència.[84] Es comprova que és necessari mantenir el nombre de neurones d'entrada per sobre de 4 (corresponent al 60% de variància explicada) per tal de garantir una capacitat de predicció acceptable.

Per tal de comprovar que la metodologia d'entrenament és capaç d'adaptar-se a diferents conjunts d'entrenament, sense renunciar a la seva capacitat de predicció, el càlcul es realitza també sobre diferents combinacions del conjunt d'entrenament. Si bé la gran majoria permet obtenir resultats equivalents als mostrats en la taula 5.4, alguns casos mostren problemes per a l'assignació de certes molècules quan es troben en el conjunt de validació. En el model només s'ha tingut en compte la permeabilitat de les molècules i no altres mecanismes d'entrada com pot ser la presència de transportadors de membrana. Aquest fet posa d'evidència que a l'hora d'establir un model de predicció, a més de tenir en compte els valors dels descriptors i de les característiques estructurals de les molècules implicades, cal també considerar els possibles processos (biològics) implicats en la propietat que es desitja predir.

Permeabilitat front a cèl·lules CACO-2

L'estudi de la permeabilitat de cèl·lules del tipus CACO-2 és, des del 1991,[87] el model cel·lular més utilitzat per a la descripció de l'absorció intestinal de fàrmacs: Aquest fet el fa especialment interessant pels estudis ADMET (d'absorció, distribució, metabolisme, excreció i toxicitat).

Per tal de comprovar que la metodologia, emprada en l'estudi de la BBB, és independent del conjunt d'entrenament quan aquest es troba correctament equilibrat, s'estudia la permeabilitat d'un conjunt de 44 fàrmacs front a cèl·lules CACO-2. A diferència de la BBB, en aquest cas existeixen models de predicció quantitatius per a la permeabilitat d'aquestes cèl·lules,[88] de manera que la seva classificació pot ser predita mitjançant el tipus de descriptors definits a l'apartat 2.1.1 (pàg. 53). Per aquest motiu es defineix com a senyals d'entrada de l'ANN les 6 primeres components principals de l'anàlisi PCA realitzada sobre el conjunt de descriptors topològics i estructurals disponibles al programari MOE.[81] En definir-se diferents conjunt d'entrenament (de 30 molècules), tots els models proposats permeten reconèixer el 100% de les molècules d'entrenament i predir el 100% de les molècules de la validació interna i externa. Si bé, el valor de reconeixement es troba dins dels valors bibliogràfics, la capacitat de predicció que mostra el model millora considerablement els trobats en la bibliografia basats en PLS.[88]

D'aquesta manera es valida el mètode d'aprenentatge i es comprova que l'equilibrat del conjunt d'entrenament no significa tan sols assegurar que tots els valors de sortida esperats s'hi trobin a parts

iguals i que totes les característiques estructurals que poden tenir les molècules a estudiar s'hi trobin representades. Cal també incloure els representats d'aquells processos que poden influir de forma indirecta el procés que s'està estudiant.

5.4 Mètodes de *pruning*

La principal idea sobre la qual es fonamenten els mètodes de *pruning* és que els pesos que no tenen gaire influència en l'etapa d'entrenament i contribueixen poc en la disminució de l'error, poden ser penalitzats i, en el límit, eliminats.[48]

Decaïment dels pesos

Aquest mètode equival a afegir una penalització dins la funció d'error emprada durant l'entrenament (eq. 5.23), que és funció del valor dels pesos que governen les connexions sinàptiques de la xarxa.

$$\mathscr{E}' = \mathscr{E} + \beta \sum_i \sum_j w_{ij}^2 \tag{5.23}$$

Això fa variar la manera com s'actualitzen els pesos (w_{ij}) durant el procés iteratiu d'aprenentatge. En el cas de l'*steepest descent* de l'algorisme *back-propagation*, cal afegir un terme corrector proporcional al valor del pes que s'està actualitzant (on $0 < \beta < 1$).[89]

$$w_{ij}(n+1) = -\eta \frac{\partial \mathscr{E}(n)}{w_{ij}} + \beta\, w_{ij}(n) \tag{5.24}$$

Amb aquesta estratègia es força que els pesos sinàptics convergeixin a valors menors, contribuint a la reducció de la variància dels valors de sortida, i aquells que no afecten al descens de l'error presenten un decaïment exponencial en el temps.[90] Tot i així, l'avaluació reiterada del valor de sortida de cadascuna de les neurones fa que aquest mètode no sigui adequat per a xarxes amb un gran nombre de neurones.[91]

Estimació de la sensibilitat sinàptica

Una altra opció consisteix en avaluar la sensibilitat de la funció error front a l'eliminació d'una determinada connexió sinàptica. En aquest cas, a cadascuna d'elles se li assigna un valor S_{ij}, que es calcula com l'evolució de l'error en presència i absència de la connexió en qüestió. Des del punt de vista dels pesos, la comparació es realitza entre l'error associat al pes sinàptic resultant del procés d'entrenament (w_{ij}^f) i el del seu valor inicial, que en la teoria es considera zero. A la pràctica, però, els mètodes d'entrenament parteixen de pesos amb un valor inicial (w_{ij}^i):

$$S_{ij} = E(w_{ij} = 0) - E(w_{ij} = w_{ij}^f) \simeq E(w_{ij} = w_{ij}^i) - E(w_{ij} = w_{ij}^f) \tag{5.25}$$

La variació de la funció error en canviar un determinat pes sinàptic (passant de w_{ij}^i a w_{ij}^f) es pot estimar com el pendent de la recta secant que passa per aquest dos punts:

$$S_{ij} = -\frac{E(w^f) - E(w^i)}{w^f - w^i} w^f \tag{5.26}$$

Tanmateix, el procés d'entrenament modifica en cada iteració tots els pesos sinàptics, de manera que tots ells contribueixen a la modificació de l'error. El fet d'incloure totes les connexions sinàptiques en la definició de la sensibilitat fa que aquesta prengui la forma:

$$S_{ij} = \int_I^F \frac{\partial E}{\partial w} d w \tag{5.27}$$

En cas d'utilitzar un entrenament basat en la regla δ, el desenvolupament de la integral anterior porta a l'expressió següent:

$$S_{ij} = \frac{1}{\eta} \left(\frac{w_{ij}}{w_{ij}^f - w_{ij}^i} \right) \sum \left(\Delta w_{ij} \right)^2 \tag{5.28}$$

Una vegada calculades, es comparen les sensibilitats de cadascuna de les connexions sinàptiques, i s'elimina aquella amb un valor mínim (que té, per tant, menys efecte en la minimització de la funció d'error). La supressió de la connexió indicada es realitza dins d'ArIS modificant el rengle corresponent de la matriu `synapse` (pàg. 84), i anul·lant els elements corresponents. La implementació final d'aquest mètode de *pruning* en el programari ArIS es realitza segons l'algorisme 7.

Algorisme 7 Mètode de pruning basat en l'estimació de la sensibilitat sinàptica.

```
1:  while CONVERGENCE=FALSE do
2:    if E ⩽ 2 · ΔE then
3:      S_ij ← w_ij: Càlcul de la sensibilitat de cada connexió
4:      p, q|S_pq = min{S_ij}
5:      w_pq ← 0: Supressió de la connexió amb menor sensibilitat
6:      Reentrenament de la xarxa
7:      if E(w_pq = 0) < E(w_pq = w_pq^f) then
8:        S'accepta el canvi
9:      else
10:       Es rebutja el canvi
11:     end if
12:   end if
13: end while
```

El mètode de *pruning* implementat en ArIS ha estat validat a nivell de descriptors moleculars, on sempre ha estat capaç de detectar les correlacions entre ells i eliminar successivament aquells menys rellevants per al model de predicció. Els resultats han comprovat l'eficàcia d'aquesta tècnica en bases de dades amb un elevat nombre de descriptors.

5.5 *Genetic Neural Networks*

El problema d'*overfitting* que pateixen molts models acostuma a ser degut a un excés d'informació en la definició del problema. La reducció del nombre de descriptors a utilitzar pot realitzar-se en una etapa posterior a l'establiment del model mitjançant el *pruning* (secció 5.4, pàg. 119), o de forma prèvia intentant trobar la aquella combinació de descriptors que permet maximitzar la capacitat de predicció.

La cerca del millor subconjunt de descriptors es pot realitzar sota el punt de vista dels mètodes d'optimització. Anteriorment s'havien implementat els algorismes genètics (GA) en el programari PRALINS, per ser utilitzats en la selecció de quimioteques combinatòries.[68] Els bons resultats obtinguts fa plantejar la seva utilització per a l'optimització dels descriptors utilitzats en una ANN.

Aquesta metodologia híbrida rep el nom de *genetic neural network* (GNN), i ha demostrat ser d'utilitat per a l'obtenció de mètodes QSAR per aplicacions de la química mèdica.[55,92–95]

5.5.1 Implementació de GNN a ArIS

Aprofitant l'arquitectura interna d'ArIS, la selecció de descriptors es pot realitzar modificant la matriu de connexions sinàptiques. La selecció dels senyals d'entrada a partir d'aquesta matriu resulta ser molt útil,[96] ja que la presència o absència d'un descriptor a la capa d'entrada de l'ANN ve determinada en últim terme per les connexions sinàptiques que estableix amb la resta de neurones. Així doncs, es pot seleccionar un conjunt de descriptors habilitant, només, les connexions sinàptiques que té associades (pàg. 84).

A l'interior dels cromosomes del GA s'hi codifiquen els identificadors dels descriptors seleccionats. Degut als problemes de convergència que s'ha pogut comprovar que presenten els cromosomes amb codificació numèrica, s'opta per una codificació binària on cada bit de la cadena correspon a un descriptor i només els descriptors seleccionats prenen el valor 1.

Donat que es desitja optimitzar la capacitat de reconeixement i predicció del model, la funció de *fitness* ha de recollir aquesta informació.[97] Per a ANN predictives es pot escollir utilitzar tant el valor de l'RMSE com el coeficient de correlació de Pearson (r). En el cas d'ANN classificatòries s'utilitza el nombre d'individus ben classificats.

El procés global del mètode GNN implementat dins ArIS correspon a l'algorisme 8.

Algorisme 8 Mètode GNN per a la selecció de descriptors.

Requeriments: Nombre total de descriptors (N). Nombre de descriptors a seleccionar (n)

1: Generar població inicial
2: Comprovació que el nombre de descriptors seleccionats sigui correcte
3: Càlcul de la funció de *fitness*
4: **while** Condició de convergència del procés evolutiu **do**
5: Selecció de dos cromosomes
6: Aplicació de l'operador d'encreuament
7: Aplicació de l'operador mutació (inclòs dins de l'encreuament)
8: Aplicació del mètode d'elitisme
9: Creació d'una nova població
10: Es recalcula el *fitness* de cada cromosoma
11: **end while**
12: Selecció del millor model

De les tècniques descrites a la introducció (secció 3.2, pàg. 67), s'ha implementat al programari ArIS l'operador de selecció *Roulette Wheel*, els operadors d'encreuament en un i dos punts (que inclouen els mecanismes de mutació corresponents) i el mètode d'elitisme.

A banda, la resta de tècniques es van implementar amb èxit al programari PRALINS, i poden ser fàcilment adaptades per ArIS en cas de considerar-se necessari.

5.5.2 Optimització de formulacions cosmètiques

L'estudi de la influència que exerceixen cadascun dels components d'una formulació sobre les seves propietats requereix realitzar un elevat nombre d'experiments. A la pràctica, aquest nombre es pot reduir aplicant un disseny d'experiències. Segons aquest esquema els membres del laboratori de Formulacions Químiques del departament d'Enginyeria Industrial de l'Institut Químic de Sarrià (IQS), estudiaren la viscositat d'un conjunt de cremes capil·lars, amb diferents proporcions dels seus components.

Davant la necessitat de disposar d'un model de predicció per a la viscositat de les formulacions estudiades (η), s'utilitzen les dades derivades del disseny d'experiències per entrenar un model de predicció quantitativa d'aquesta propietat. Els resultats obtinguts foren posteriorment publicats,[98] veure annex IV.

L'entrenament es realitza mitjançant l'algorisme BP-Batch, definint-li una validació interna del tipus LOO. El millor resultat obtingut correspon a una ANN de topologia 8-5-1 que presenta una $R^2 = 0.99$ per al conjunt d'entrenament i una $R^2 = 0.85$ per al conjunt de validació externa, figura 5.10.

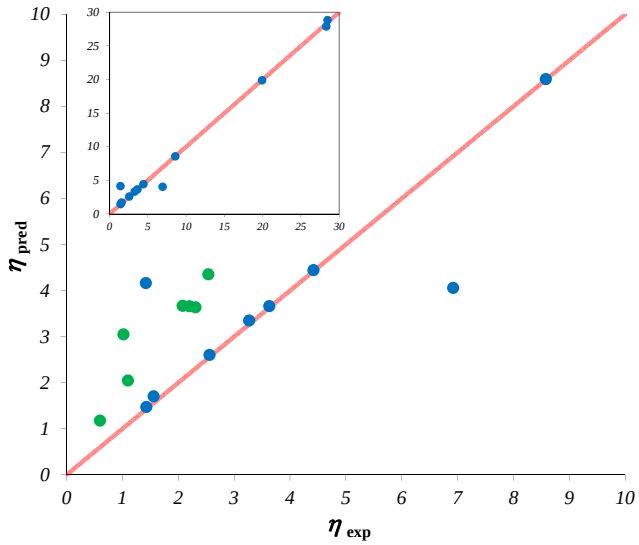


Figura 5.10: *Resultats obtinguts per a la predicció amb ANN de la viscositat de formulacions químiques. [• conjunt d'entrenament, • conjunt de validació]*

L'algorisme GNN s'utilitza per identificar els descriptors amb menor influència sobre la definició de la viscositat. Així, es permet passar a una ANN de topologia 6-4-1, en la qual s'eliminen dues de les components de les formulacions. El resultat final (figura 5.11), permet reduir el RMSE del conjunt de validació de 0.55 (de la topologia 8-5-1) a un valor de 0.23.

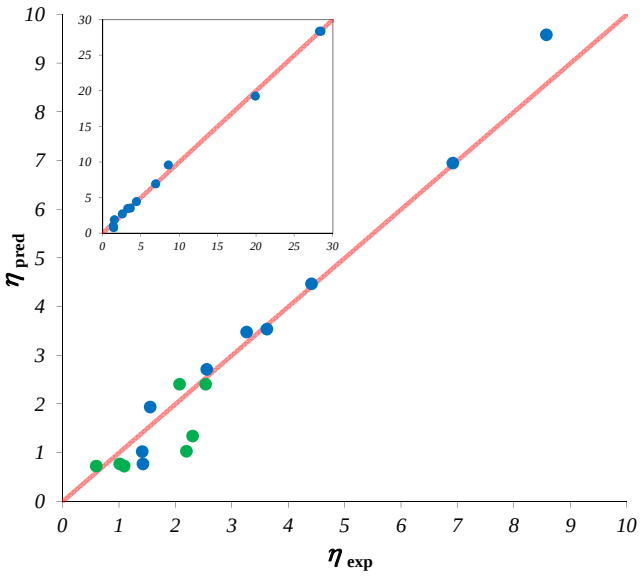


Figura 5.11: *Resultats obtinguts per a la predicció de la viscositat de formulacions químiques després d'aplicar l'algorisme GNN per a la selecció de descriptors. [• conjunt d'entrenament, • conjunt de validació]*

5.5.3 Estudi complementari dels anàlegs d'AZT

El mètode GNN implementat en ArIS també s'ha aplicat per avaluar la reducció del nombre de descriptors en l'estudi realitzat sobre els anàlegs de l'AZT (pàg. 115). Els autors de l'article pres de referència posen de manifest la inconveniència de reduir a quatre el nombre de descriptors, mitjançant un mètode de *pruning*. Per aquest motiu, s'estudia l'efecte d'eliminar només un descriptor, passant de 6 a 5 amb GNN.

Els resultats obtinguts permeten identificar el nombre de Wiener com a possible descriptor a eliminar. La xarxa de topologia 5-6-10-2 obtinguda permet mantenir els resultats inicials amb 6 descriptors (100% de reconeixement de les dades d'entrenament i 86% (12/14) en la predicció del conjunt de validació).

5.6 Resum

S'han implementat i validat els mètodes corresponents a perceptrons, combinadors lineals adaptatius d'una o més neurones i les xarxes neuronals multicapa amb l'entrenament de retropropagació de l'error.

Complementàriament, s'han incorporat els mètode de *pruning* i GNN per la identificació dels descriptors més rellevants i facilitar la sreducció del nombre de senyals d'entrada sense afectar significativament el resultat obtingut.

La utilització de xarxes amb més d'una capa permet s'estudi de sistemes complexos, demostrant millors resultats que els mètodes lineals.

Capítol 6

Reflexions sobre la representació gràfica dels resultats

Els mètodes computacionals utilitzen, sovint, més de dos descriptors moleculars, que juntament amb la funció resposta fan que la definició matemàtica del problema a tractar correspongui a un espai n-dimensional. Si bé els mètodes quimioinformàtics poden treballar sense problemes en més d'una dimensió, la discussió i interpretació gràfica d'aquestes dades esdevé un problema quan se supera la tercera dimensió.

Per aquest motiu, s'estudia la manera d'incorporar a ArIS una eina de visualització d'espais n-dimensionals que serveixi per inspeccionar la distribució inicial d'unes determinades dades o per avaluar l'ús de diferents descriptors. Donat que la matemàtica és especialista en treballar amb espais multidimesionals, es comença cercant resposta en la geometria.

6.1 Mètodes per a la reducció de la dimensió de l'espai

En el camp de la quimioinformàtica ja es disposa d'un conjunt d'estratègies, suficientment validades, per reduir la dimensionalitat. L'anàlisi discriminant (*Discriminant Analysis*, DA) i l'anàlisi de components principals (*Principal Components Analysis*, PCA), són dues de les tècniques més utilitzades actualment. Aquesta última, es troba fins i tot implementada en la majoria de programaris destinats al disseny molecular.

L'anàlisi de components principals

L'anàlisi de components principals permet passar d'un conjunt de variables (que poden estar correlacionades) a un conjunt menor de variables ortogonals, anomenades components principals, que permeten explicar la mateixa variància que les originals. Els nous eixos de coordenades es defineixen com una combinació lineal de les variables originals.

Sigui $\{p_i\}$ un conjunt de variables (mesurades o calculades computacionalment), amb les quals

es realitzen n observacions. La informació obtinguda es pot expressar en forma matricial com:

$$\mathbb{Y} = \begin{pmatrix} p_1(1) & p_2(1) & \cdots & p_i(1) \\ p_1(2) & p_2(2) & \cdots & p_i(2) \\ \vdots & \vdots & \ddots & \vdots \\ p_1(n) & p_2(n) & \cdots & p_i(n) \end{pmatrix} \tag{6.1}$$

Normalment les dades se centren segon la seva mitjana i es normalitzen, donant lloc a la matriu $\mathbb{X}$. A continuació es calcula la seva matriu de variància-covariància com el producte $\mathbb{Z} = \mathbb{X}^T \cdot \mathbb{X}$, on

$$\mathbb{X} = \begin{pmatrix} \frac{y_{11}-\overline{y_{j1}}}{\sigma_{y_{j1}}} & \frac{y_{12}-\overline{y_{j2}}}{\sigma_{y_{j2}}} & \cdots & \frac{y_{1n}-\overline{y_{jn}}}{\sigma_{y_{jn}}} \\ \frac{y_{21}-\overline{y_{j1}}}{\sigma_{y_{j1}}} & \frac{y_{22}-\overline{y_{j2}}}{\sigma_{y_{j2}}} & \cdots & \frac{y_{2n}-\overline{y_{jn}}}{\sigma_{y_{jn}}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{y_{n1}-\overline{y_{j1}}}{\sigma_{y_{j1}}} & \frac{y_{n2}-\overline{y_{j2}}}{\sigma_{y_{j2}}} & \cdots & \frac{y_{nn}-\overline{y_{jn}}}{\sigma_{y_{jn}}} \end{pmatrix} \tag{6.2}$$

La diagonalització de la matriu $\mathbb{Z}$ permet calcular-ne els valors (VAPS) i vectors propis (VEPS). Els VAPS informen del percentatge de variància que permet explicar en l'eix definit per la direcció del VEP corresponent (corresponent a la component principal). Així doncs, l'eix principal vindrà definit per aquell VAP de valor major. A aquest se li pot afegir les contribucions d'altres VAPS, podent disposar d'un elevat control sobre el percentatge de la variància explicada. Per exemple, si hom desitja explicar el 90% de la variància de les dades originals, haurà de seleccionar aquell conjunt de q components principals tals que els seus VAPs (λ_i) compleixin:

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_q}{\sum \lambda_i} \geqslant 0.9 \tag{6.3}$$

Tot i que normalment és suficient amb un nombre reduït de components principals per explicar una proporció significativa de la variància de les dades, pot ser que aquest continuï sent superior a 3. En escapar-se de l'espai tridimensional, aquest casos continuen sense poder-se representar gràficament.

6.2 L'abisme d'Euclides

Un dels objectius de l'educació primària a Catalunya és, segons el decret 142/2007, de 26 de juny (capítol 1, article 3),[99] contribuir a *desenvolupar les competències matemàtiques bàsiques, iniciar-se en la resolució de problemes que requereixin la realització d'operacions elementals de càlcul,* ***coneixements geomètrics i estimacions, i ser capaç d'aplicar-les a les situacions de la vida quotidiana.***

Des de ben petits, doncs, hem estat iniciats en la geometria i hem après a relacionar-la amb els elements del nostre entorn. Aquest aprenentatge es va desenvolupant, poc a poc i al llarg dels diferents cursos acadèmics, fins arribar al gran apogeu de la geometria en l'espai tridimensional. Tot aquest extraordinari camí per les rectes, els plans, les hipèrboles, els cilindres, etc. es realitza nor-

malment sota el punt de vista de tres eixos perpendiculars x, y i z, regits per les propietats de l'espai euclidià.

Una vegada apresa, la geometria euclidiana (anomenada també geometria parabòlica) ens permet modelar i estudiar la major part de les situacions amb què ens podem trobar a la vida quotidiana, sense renunciar a una certa simplicitat matemàtica (no és per casualitat que es prengui de referència). Aquesta afable situació pot excavar, emperò, un immens abisme que dificulta el salt a altres formulismes geomètrics.

La geometria euclidiana es basa en cinc postulats, recollits en els tretze llibres que conformen l'obra *Els Elements*, escrita per Euclides l'any 300 A.C.:[100]

1. Dos punts de l'espai defineixen una recta

2. Qualsevol segment rectilini pot ésser estès indefinidament formant una recta

3. Donat un segment rectilini qualsevol i un punt (P) en el pla, es pot dibuixar un cercle prenent el segment com a radi i el punt com a centre.

4. Tots els angles rectes són congruents.

5. Donades dues rectes que es tallen amb una tercera, si la suma dels angles interns d'un dels costats és inferior a dos angles rectes, aleshores les dues rectes es tallaran inevitablement si s'estenen suficientment en l'espai.

El darrer d'aquests postulats, conegut també com el postulat paral·lel d'Euclides, no s'ha pogut demostrar com a teorema i ha obert la porta a la geometria no euclidiana. Aquesta 'nova' geometria deixa de banda la planarietat aportada per la geometria parabòlica i proposa els seus propis axiomes. Per a la finalitat d'aquest capítol és suficient considerar aquelles geometries amb una curvatura negativa constant, les quals accepten els quatre primers postulats, però utilitzen la seva pròpia definició del postulat paral·lel: la geometria de Riemann (o geometria el·líptica) i la geometria hiperbòlica.[101]

6.2.1 La geometria hiperbòlica

El postulat paral·lel hiperbòlic estableix que donada una recta l i un punt $\mathscr{P}$ qualsevol no pertanyent a l, existeixen múltiples rectes que passen per $\mathscr{P}$ i són paral·leles a l.[102]

Per tal de poder representar gràficament aquest tipus de geometria, s'han desenvolupat diferents models que permeten projectar l'espai hiperbòlic sobre el pla euclidià. En aquest apartat només es considerarà el disc de Poincaré. Aquestes tècniques s'han aplicat amb èxit per la representació gràfica de xarxes n-dimensionals.[103–105]

En efectuar aquestes aplicacions sobre l'espai hiperbòlic, per obtenir la seva projecció sobre el pla euclidià, es podria veure compromesa la correcta representació de la distància entre els punts, les àrees, els angles, etc. Per aquest motiu cal aplicar unes operacions matemàtiques, anomenades isometries, que garanteixin el manteniment de les distàncies entre espais mètrics diferents. En el cas dels models anteriors, les isometries venen donades per **transformacions de Möebius**.[106]

Tal com es pot entreveure, aquest fet és en part degut al canvi de mètrica que es produeix entre els espais de sortida i arribada. No es considera adequat en aquest punt entrar en la descripció de l'espai euclidià d'arribada, apel·lant a l'extens coneixement que es té d'ell. En canvi, l'espai hiperbòlic de sortida és un espai amb una curvatura negativa constant, de manera que les eines per descriure les corbes i les superfícies en ell haurà de ser diferent que el cas anterior. En aquest sentit, la **mètrica de Poincaré** és un tensor mètric que descriu una superfície bidimensional amb una curvatura negativa constant, sent la mètrica utilitzada normalment per descriure l'espai hiperbòlic. Un tensor mètric correspon a una entitat algebraica que, definida sobre una superfície, permet generalitzar la definició del producte escalar i ésser aplicat en la definició de distàncies, angles, etc. de forma independent al sistema de coordenades emprat.[107]

Primera forma quadràtica fonamental

Es considera $\mathscr{U}$ un conjunt obert de $\mathbb{R}^2$ i la funció $f : \mathscr{U} \to \mathbb{R}^3$. Sigui $\mathscr{S}$ la superfície definida per $f(\mathscr{U})$, $u_0 \in \mathscr{U}$ i $t \to u(t)$ una corba diferenciable de $\mathscr{U}$ que passa per u_0.

Es defineix $T_{f(u)}(\mathscr{S})$ com l'espai tangent a la superfície $\mathscr{S}$ en el punt $f(u_0)$, i corresponent al subespai de $\mathbb{R}^3$ format pel conjunt dels vectors tangents a les corbes $u(t)$ en aquest punt ($t = 0$). Tot vector tangent a la superfície $\mathscr{S}$ pot ser expressat com a combinació lineal de les derivades parcials de la corba respecte les coordenades de l'espai $\mathbb{R}^2$, esdevenint base de l'espai $T_{f(u)}(\mathscr{S})$.

Definit el producte escalar (g) sobre l'espai tangent a una superfície, donat que $\{\frac{\partial f}{\partial u^1}, \frac{\partial f}{\partial u^2}\}$ és base de l'espai, el resultat del producte escalar pot expressar-se com a combinació lineal del producte escalar dels vectors de base:

$$\mathbb{G} = \begin{pmatrix} E & F \\ F & G \end{pmatrix} \tag{6.4}$$

$$\text{on } E = g\left(\frac{\partial f}{\partial u^1}, \frac{\partial f}{\partial u^1}\right), F = g\left(\frac{\partial f}{\partial u^1}, \frac{\partial f}{\partial u^2}\right), G = g\left(\frac{\partial f}{\partial u^2}, \frac{\partial f}{\partial u^2}\right).$$

El producte escalar $g(T_x(\mathscr{S}), T_x(\mathscr{S}))$ rep el nom de **forma quadràtica fonamental** d'una superfície i esdevé clau en la definició del tensor mètric que regeix l'espai.

Definició de la longitud entre dos punts

Sigui la superfície $\mathscr{S}$ parametritzada sobre $\mathbb{R}^3$ i generada a partir d'una funció $f : \mathscr{U} \to (S)$. Sigui la corba $u(t)$ una corba de $\mathscr{U}$ definida per $x(t) = f(u(t))$. La longitud de la corba en un interval determinat es pot definir com la integral de línia corresponent a l'equació 6.5:

$$L = \int_C \|\dot{x}(t)\| \, dt \tag{6.5}$$

Per la definició de norma, l'equació anterior es pot reescriure en termes del producte intern donat que $\|\dot{x}(t)\| = \sqrt{\langle \dot{x} | \dot{x} \rangle}$. D'aquesta manera, si es desenvolupa el producte intern d'acord amb la definició exposada anteriorment i considerant que $\mathscr{U} \in \mathbb{R}^2$, hom obté:

$$L=\int_C \sqrt{\sum_{i,j}\left\langle \frac{\partial f}{\partial u^1}|\frac{\partial f}{\partial u^2}\right\rangle \frac{\partial u^i}{\partial t}\frac{\partial u^j}{\partial t}}\,dt=\int_C \sqrt{\sum_{i,j} g_{i,j}\frac{\partial u^i}{\partial t}\frac{\partial u^j}{\partial t}}\,dt$$

$$L=\int_C \sqrt{E\left(\frac{\partial u^1}{dt}\right)^2+2F\frac{\partial u^1}{dt}\frac{\partial u^2}{dt}+G\left(\frac{\partial u^2}{dt}\right)^2}\,dt \tag{6.6}$$

La definició de les geodèsiques

La redefinició del postulat paral·lel fa que tot i saber com mesurar la distància entre dos punts d'un espai amb mètrica hiperbòlica, cal definir el camí per on mesurar-la donat que no té per què ser una recta. Així doncs, l'expressió del càlcul d'L es pot interpretar com una funció que, aplicada sobre una segona funció, retorna un valor real (és a dir, es pot considerar com un funcional). La funció d'entrada que permet minimitzar el valor de la distància entre dos punts defineix les geodèsiques de l'espai.

Sigui $\alpha(t)$ la corba parametritzada per arc que minimitza la distància entre dos punts P i Q, corresponents als valors $t=0$ i $t=a$ respectivament. Es considera una nova funció h, que depèn de la variable $s\in\mathbb{R}$, i que permet definir una variació de la corba anterior, diferenciable, i amb extrems fixos P i Q:

$$\alpha(s,t)=\alpha(t)+sh(t)$$

Així, totes les possibles corbes que uneixen ambdós punts poden expressar-se com la variació $\alpha(s,t)$ (figura 6.1), on $\alpha(0,t)=\alpha(t)\forall t\in[0,a]$.

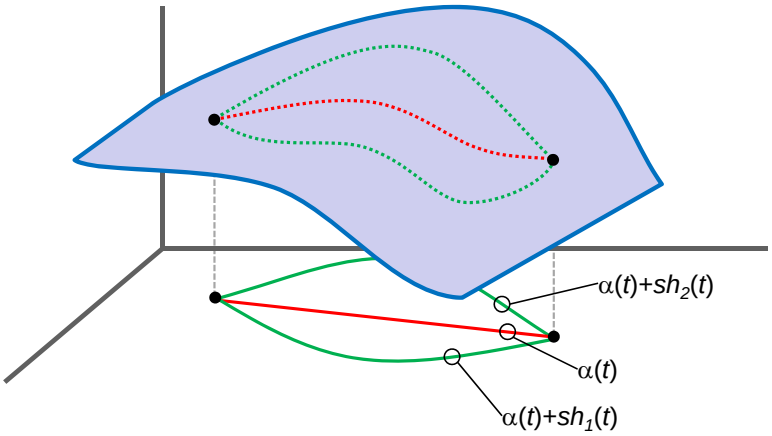


Figura 6.1: *Representació esquemàtica d'una variació de la corba $\alpha(t)$ entre dos punts.*

Per la condició de mínim, doncs, cal que complexi:

$$\left(\frac{d}{ds}L(\alpha)\right)_{s=0}=0 \tag{6.7}$$

En trobar-se parametritzada per arc, la corba $\alpha(t)$ presenta una norma unitat en aplicar-li l'operador $\frac{\partial}{\partial t}$, fet que permet reescriure l'equació anterior com:

$$\left(\frac{d}{ds}L(\alpha)\right)_{s=0} = \frac{1}{2}\int_0^a \frac{\partial}{\partial s} g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right)_{s=0} dt \tag{6.8}$$

Utilitzant les propietats de la derivada covariant sobre l'eq. 6.8, es pot demostrar[106] que la variació del producte intern respecte de la variable s es pot expressar en forma d'un sol producte intern que impliqui aquest tipus d'operació:

$$\left(\frac{d}{ds}L(\alpha)\right)_{s=0} = \int_0^a g\left(\frac{\partial}{\partial s}, \nabla_{\frac{\partial}{\partial t}}\frac{\partial}{\partial t}\right)_{s=0} dt = 0 \tag{6.9}$$

Si aquesta igualtat s'ha de complir per qualsevol variació $\alpha(s,t)$, significa que la corba parametritzada $\alpha(t)$ serà extremal de la longitud sempre i quan $\nabla_{\frac{\partial}{\partial t}}\frac{\partial}{\partial t} = 0$, és a dir, quan $\nabla_{\dot{x}(t)}\dot{x}(t) = 0$.

El concepte de derivada covariant (d'un camp Y en la direcció X, $\nabla_X Y$) és una generalització de la derivada direccional, que permet calcular la component tangent de la derivada en una superfície[106] (és a dir, la derivada al llarg d'un vector tangent a la superfície). Cal pensar que en estar sobre una superfície definida per unes coordenades corbes, la consideració de la derivada direccional deixa de ser vàlida, ja que els vectors tangents a les línies coordenades emprades com a base (e_j) canvien en passar d'un punt a l'altre de la corba. Això comporta que s'hagi d'avaluar la variació dels vectors de base a mesura que es recorre la corba de coordenades:

$$\nabla_X Y = \sum_{i=1}^{n} \frac{\partial y^i}{\partial x^j} e_k + \sum_{i=1}^{n} y^i \nabla_X e_j \tag{6.10}$$

L'eq. 6.10 mostra com la derivada direccional (primer sumand) és modificada per la correcció comentada anteriorment. La manera com es pot definir una relació entre la geometria local d'un punt amb la d'un altre, es realitza per mitjà dels símbols de Christoffel (Γ_{ij}^k):

$$\nabla_X e_j = \sum_{m=1}^{n} \Gamma_{ji}^m e_m \tag{6.11}$$

Amb la incorporació d'aquests símbols dins l'eq. 6.10, i emprant la notació d'Einstein (s'obvien els sumatoris, els índexos dels quals no presenten confusió), hom pot expressar la derivada covariant:

$$\nabla_X Y = x(y^j)e^j + y^j x^i \Gamma_{ij}^k e_k \tag{6.12}$$

Aquest resultat significa que les geodèsiques de l'espai hauran de complir, per a cadascuna de les variables (j), el sistema d'equacions diferencials següent:

$$\nabla_{\dot{x}}\dot{x} = \dot{x}(\dot{x})^t e^j + \dot{x}^j \dot{x}^i \Gamma_{ij}^k e_k = 0 \tag{6.13}$$

és a dir

$$\frac{d^2x^j}{dt^2} + \Gamma_{ij}^k \frac{dx^j}{dt}\frac{dx^k}{dt} = 0 \tag{6.14}$$

6.3 El disc de Poincaré

El model d'espai hiperbòlic aportat pel disc de Poincaré permet representar el pla euclidià dins un disc de radi finit i unitari; $\mathscr{H}^2 = \{z \in \mathbb{C} : |z| < 1\}$.

6.3.1 La mètrica de Poincaré

La mètrica de Poincaré permet descriure una superfície bidimensional amb una curvatura negativa constant. Tot i que pot ser definida per un espai *n*-dimensional, donat que l'objectiu de l'estudi és la projecció sobre el pla, es considera la seva definició bidimensional:

$$ds^2 = \frac{dx^2 dy^2}{\left(1-\left(x^2+y^2\right)\right)^2} \tag{6.15}$$

En trobar-se definit sobre el camp dels nombres complexos, hom pot emprar la notació polar per expressar la mètrica:

$$ds = \frac{2dr}{1-r^2} \tag{6.16}$$

Segons l'eq. 6.6, la definició de la mètrica permet calcular la distància entre dos punts d'aquest espai. Si es pren com a punt de partida el càlcul de la distància entre un punt del disc i el centre, aquesta es pot calcular segons l'eq. 6.17:

$$L = \int_0^r \frac{2}{1-r^2} dr = \tanh^{-1}(r) \tag{6.17}$$

6.3.2 Les transformacions de Möebius

En representar gràficament dades en l'espai hom es troba amb la necessitat de manipular els punts, realitzant rotacions o translacions, sigui per facilitar-ne la seva interpretació o bé per millorar la seva aparença. Aquests procediments, trivials en el cas de la geometria euclidiana, requereixen d'operacions un tant més complexes quan es considera una geometria no euclidiana. Les transformacions de Möebius corresponen als operadors rotació i translació en el pla hiperbòlic $\mathscr{H}^2$.

Tota transformació conformal (i.e. que el seu resultat preserva l'orientació dels angles inicials) d'$\mathscr{H}^2$ es pot descompondre en un terme de rotació (eq. 6.18) i un operador que correspon a l'anàleg d'una translació (eq. 6.19).[108]

$$g_\tau = e^{i\tau} z \tag{6.18}$$

$$\phi_a(z) = \frac{z-a}{1-\bar{a}z} \tag{6.19}$$

Aquestes operacions permeten canviar la orientació dels punts al llarg de tot el pla complex i fins i tot centrar-lo en un punt concret (figura 6.2).

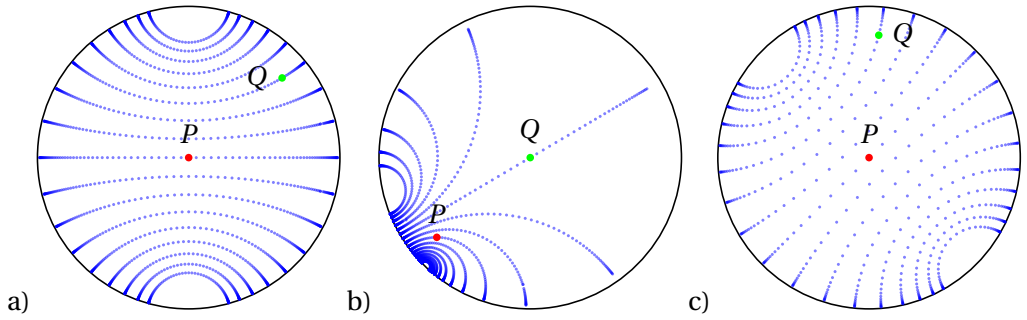


Figura 6.2: *Representació de les geodèsiques que regeixen el disc de Poincaré (a). Prenent de referència els punts anteriors: resultat de l'aplicació de les transformacions de Möebius en la translació del punt Q a l'origen (b) i aplicació de l'operador rotació (eq.6.18, $\tau = \frac{\pi}{4}$) (c).*

Tenint en compte les transformacions anteriors, el càlcul de la distància entre dos punts qualssevol de l'espai $\mathscr{H}^2$ es redueix al càlcul de la distància entre un punt i el centre del disc (eq. 6.17). Donats dos punts, es realitza en primer lloc la translació del centre del disc sobre el primer punt, i a continuació es mesura la distància amb el segon punt com la longitud entre les noves coordenades d'aquest i el centre del disc. El procediment global es recull dins d'eq. 6.20.

$$L = 2\tanh^{-1}\left(\frac{|z_i z_j|}{|1 - z_i \bar{z}_j|}\right) \tag{6.20}$$

6.3.3 Les geodèsiques

La resolució del sistema d'equacions diferencials (eq. 6.14) associat a la mètrica de Poincaré, permet identificar les trajectòries que minimitzen la distància entre dos punts de l'espai hiperbòlic. Les geodèsiques del model del disc de Poincaré descriuen arcs circulars, els extrems dels quals són perpendiculars al contorn del disc.

Numèricament, les funcions parametritzades que esdevenen solució del sistema d'equacions diferencials anterior corresponen a aplicacions de dues variables tals que, aplicades sobre un punt del pla euclidià XY, retornen la seva imatge en el disc $\mathscr{H}^2$:

$$\begin{cases} f : \mathbb{R}^2 & \rightarrow & \mathbb{C} \\ (x,y) & \rightarrow & f(x,y) = a + bi \end{cases}$$

$$f(x,y) = \left(\frac{\sinh(x)}{1+\cosh(x)\cosh(y)}\right) + \left(\frac{\cosh(x)\sinh(y)}{1+\cosh(x)\cosh(y)}\right) i \tag{6.21}$$

La figura 6.2 inclou la representació gràfica d'un conjunt d'aquestes geodèsiques, juntament amb les transformacions que pateixen en aplicar les transformacions de Möebius.

Tanmateix, la representació gràfica de la geodèsica que defineix la distància entre dos punts concrets d'$\mathscr{H}^2$ es pot realitzar geomètricament, com l'arc d'una circumferència que passa per ambdós punts i és perpendicular al disc, figura 6.3.

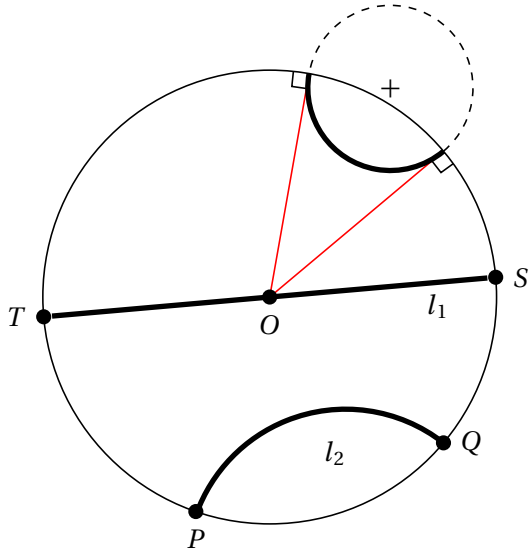


Figura 6.3: *Representació gràfica del model de disc de Poincaré. Es mostren dues rectes paral·leles: l_1, que passa pels punts T i S i recorda a una recta euclidiana, i l_2, corresponent a la geodèsica que uneix P amb Q. Tal com es pot observar en la part superior de la il·lustració, la construcció dels arcs que uneixen dos punts en l'espai hiperbòlic es realitza considerant la circumferència que talla perpendicularment els límits del disc.*

6.4 Optimització de la projecció sobre $\mathscr{H}^2$

En fer encabir tot el pla euclidià en un disc de radi finit, el model del disc de Poincaré permet visualitzar d'una forma fàcil i entenedora punts situats a llarga distància en l'espai euclidià.

Es considera l'espai d'N descriptors, associat a una determinada quimioteca, com un conjunt de punts N-dimensionals on cadascun d'ells dóna informació d'una molècula en concret. Aquest espai $\mathbb{R}^N$ es pot representar gràficament sobre el pla $\mathscr{H}^2$. Per realitzar-ho, se cerquen aquelles coordenades del disc de Poincaré on la distància entre tots els punts implicats, permetin mantenir la seva dissimilitud.

Per evitar confusions en la nomenclatura, es considera el conjunt de punts $a_1, a_2, \ldots, a_k \in \mathbb{R}^N$ com els punts corresponents a la representació molecular en l'espai euclidià; mentre que el conjunt $x_1, x_2, \ldots, x_k \in \mathbb{C}$ correspon a les respectives projeccions sobre el pla hiperbòlic.

Mesura de la semblança molecular

La projecció sobre el pla hiperbòlic requereix adoptar una definició de la semblança molecular. Tot i que aquesta qüestió podria semblar quelcom trivial des del punt de vista de l'expertesa d'un professional de l'àmbit químic, no ho és des del punt de vista computacional, doncs la definició de semblança que es cerca ha de ser intel·ligible per un ordinador.

Normalment es considera la disposició de les molècules dins l'espai de descriptors i es mesura la seva semblança d'acord amb la distància que les separa dins de l'espai N-dimensional. Així, la dissimilitud entre dos vectors (D_{ij}), és a dir, la manera com es defineix quant diferents són una parella de punts, es pot mesurar de diferents maneres segons la mètrica emprada. A més de la definició

euclidiana, en la qual es defineix la distancia com el mòdul del vector que uneix a dos punts (eq. 6.22), en aquest apartat també es considera la definició de la semblança basada en el cosinus de l'angle que formen dos punts (eq. 6.23) i la mètrica de Canberra (eq. 6.24).[109]

$$D_{ij} = \|a_i - a_j\| = \sqrt{\left(a_i^1 - a_j^1\right)^2 + \ldots + \left(a_i^N - a_j^N\right)^2} \tag{6.22}$$

$$D_{ij} = 1 - \frac{\bar{a}_i \bar{a}_j}{\|a_i\|\|a_j\|} \tag{6.23}$$

$$D_{ij} = \sum_{k=1}^{N} \frac{|a_{ik} - a_{jk}|}{\|a_{ik}\|\|a_{jk}\|} \tag{6.24}$$

La definició d'un criteri de semblança sobre un espai de descriptors determinat, permet fer l'extrapolació de la matemàtica a la química, considerant que molècules semblants a nivell de descriptors presentaran propietats químiques similars (segons el principi de semblança de Johnson i Maggiora[110]).

6.4.1 Mètodes d'optimització considerats

Una vegada es disposa de la definició de semblança molecular, l'objectiu de la projecció sobre $\mathscr{H}^2$ és poder representar tot l'espai de descriptors dins d'un disc de radi finit, garantint que es mantingui la dissimilitud entre els punts. Per aquest motiu es requereixen mètodes d'optimització que permetin trobar les coordenades $\{x_i\}$ que compleixin la condició anterior.

Steepest Descent

El mètode *Steepest Descent* és un mètode relativament estès per a la realització d'optimitzacions.[111] Segons aquest, l'actualització de les coordenades de la projecció es modifiquen de forma iterativa seguint l'eq. 6.25, aprofitant la informació de la direcció del gradient en un punt (Δx).

$$x_i(t+1) = x_i(t) + \eta \Delta x \tag{6.25}$$

Donat que x_i correspon a un punt del disc de Poincaré, la translació d'aquest punt una certa magnitud, $\eta\Delta x$, no es pot realitzar de forma directa. Tal com s'ha comentat en l'apartat 6.3.2, cal aplicar la transformació de Möebius corresponent a aquesta operació. Això significa que la definició del punt següent indicada en l'eq. 6.25 passa a prendre la forma següent:

$$x_i(t+1) = T\left(x_i(t), \eta\Delta x, 1\right) \tag{6.26}$$

on $T(z,c,\theta) = \frac{\theta z + c}{1 + \theta z \bar{c}}$ correspon a una transformació de Möebius que descriu la translació dins de la projecció hiperbòlica.

Sent el gradient el vector indicador de la direcció de màxim pendent, el seu seguiment permet conduir el valor d'una funció al punt estacionari més proper. En aquest cas, la funció que es desitja minimitzar ha de recollir la discrepància entre la distància hiperbòlica entre dos punts (d_{ij}, calculada

mitjançant l'eq. 6.17) i la seva dissimilitud (D_{ij}). Per garantir la dissimilitud en la projecció hiperbòlica s'empra normalment la funció de Sammon[111] (eq.6.27).

$$E(x_{ij}) = \sum_{i=1}^{N} \sum_{j>i}^{N} w_{ij} \left(d_{ij} - D_{ij} \right)^2 \tag{6.27}$$

Es considera, doncs, que la funció de Sammon reflecteix l'error comès en la representació bidimesional sobre $\mathscr{H}^2$, de manera que per minimitzar-la cal calcular el gradient associat a ella. Per aquesta finalitat s'avalua la primera derivada de la funció error respecte les coordenades hiperbòliques.

$$\frac{\partial E(x_{ij})}{\partial x_i} = \frac{\partial E}{\partial d_{ij}} \frac{\partial d_{ij}}{\partial x_i} \tag{6.28}$$

$$\frac{\partial E(x_{ij})}{\partial x_i} = \sum_{j>i}^{N} 2 w_{ij} \left(d_{ij} - D_{ij} \right) \frac{\partial d_{ij}}{\partial x_i} + \sum_{j=1}^{i-1} 2 w_{ji} \left(d_{ji} - D_{ji} \right) \frac{\partial d_{ji}}{\partial x_i} \tag{6.29}$$

Donat que l'optimització es realitza sobre l'espai definit pel disc de Poincaré, cada coordenada està formada per una component real i una d'imaginària, de manera que:

$$\frac{\partial d_{ij}}{\partial x_i} = \frac{\partial d_{ij}}{\partial Re(x_i)} + \left(\frac{\partial d_{ij}}{\partial Im(x_i)} \right) i \tag{6.30}$$

La substitució d'aquestes derivades en la definició de l'algorisme d'optimització *Steepest Descent* modificat, permet recalcular de forma iterativa les projeccions de cadascun dels punts, minimitzant de forma progressiva la funció error escollida.

Aplicació dels algorismes genètics

En un treball previ s'havien implementat els algorismes genètics en el programari PRALINS.[68] Aprofitant aquest coneixement i de forma independent, es proposa l'ús dels algorismes genètics (secció 3.2, pàg. 67) com a alternativa als mètodes d'optimització descrits bibliogràficament,[103] basats en la informació del gradient.

Els algorismes genètics són mètodes d'optimització que permeten explorar de forma ràpida i eficaç la totalitat de l'espai de respostes. El conjunt de coordenades que conformen una resposta per a la projecció hiperbòlica rep el nom de cromosoma i és assimilat a un individuu d'una població biològica. Com a tal, aquesta població evoluciona amb el temps seguint les lleis naturals de supervivència del més ben adaptat a l'entorn, i.e. la solució que permet reduir més el valor de la funció de l'error proposta per Sammon (que en aquest context rep el nom de funció de *fitness*).

L'evolució biològica se simula per mitjà de tres operadors, que s'apliquen en aquest ordre:

1. Selecció: Dos elements de la població són seleccionats mitjançant l'algorisme *Roulette Wheel*, on la probabilitat de selecció és inversament proporcional a la seva funció de *fitness*.

2. Encreuament: La informació codificada en els cromosomes seleccionats es barreja intercanviant els gens (*i.e.* les coordenades) a partir d'un punt determinat aleatòriament.

3. Mutació: Els errors produïts durant la transcripció i traducció del material genètic se simulen per l'alteració de gens seleccionats aleatòriament després de l'encreuament. Aquest operador facilita l'exploració de l'espai de respostes i evita la homogeneïtzació de la població a un cromosoma dominant.

6.4.2 Implementació d'un programari específic

S'ha creat un programari, en llenguatge Visual Basic mitjançant Visual Studio 2010,[76] que permet realitzar la projecció d'un conjunt de dades n-dimensionals sobre el disc de Poincaré. Tant la seva interfície com la implementació algorísmica s'han dissenyat per fer-lo genèric i adaptable a qualsevol aplicació: algorisme 9.

Algorisme 9 Projecció sobre el pla hiperbòlic $\mathscr{H}^2$.

Requeriments: Nombre màxim d'iteracions (t_{max}), llindar

Retorna: $\{x_i\} \leftarrow \{a_i\}$

1: Lectura de dades a_i
2: $D_{ij} \leftarrow$ Càlcul de les dissimilituds
3: $w_{ij} \leftarrow$ Càlcul del factor de normalització
4: $x_i(0) \leftarrow$ Establiment de les coordenades inicials
5: $E(0) \leftarrow$ Càlcul de la funció de *fitness*
6: **while** ($t < t_{max}$ **o** $E(t) >$ llindar) **do**
7: $\quad t \leftarrow t+1$
8: $\quad x_i(t) \leftarrow$ Actualitzar les coordenades de la projecció
9: $\quad E(t) \leftarrow$ Recalcular la funció de *fitness*
10: **end while**
11: Representar gràficament

En l'algorisme 9 es pot veure com l'estructura de repetició, delimitada per dos criteris de convergència (número d'iteracions màximes i el valor llindar per sota del qual es considera que l'error és assumible), correspon al mètode d'optimització que recalcula de forma iterativa un millor conjunt de coordenades per a les projeccions dels punts inicials.

El programa desenvolupat disposa tant del mètode *Steepest Descent* com d'un algorisme genètic per a l'optimització de les coordenades. En ambdós casos la semblança entre dos punts (x i y), entesa com la distància D que els separa, pot definir-se mitjançant:

1. La distància euclidiana:

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_L - y_L)^2} \tag{6.31}$$

2. La mètrica del cosinus:[103]

$$D = 1 - \frac{\overline{x} \cdot \overline{y}}{|x| \cdot |y|} \tag{6.32}$$

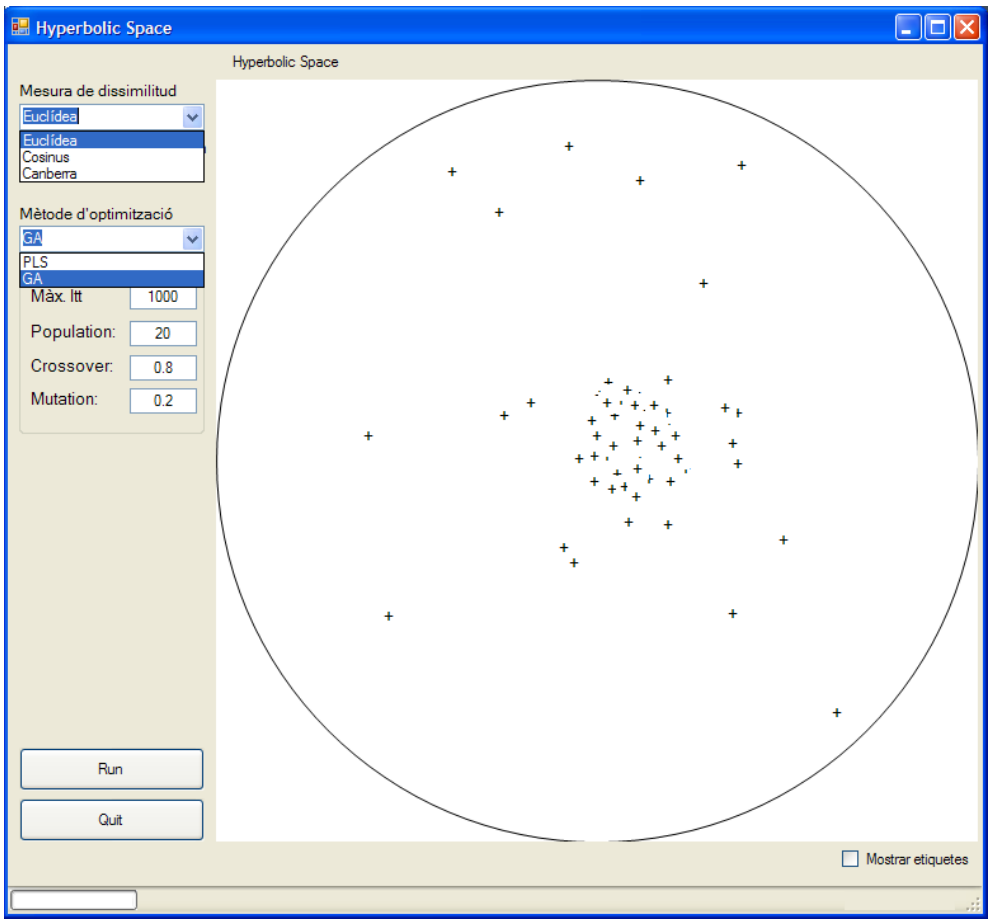
3. La mètrica de Canberra:[109]

$$D = \sum_{k=1}^{L} \frac{|x-y|}{|x|+|y|} \tag{6.33}$$

Amb la finalitat de millorar la representació gràfica de les dades dins del disc de Poincaré i ajudar a la utilització de tot l'espai disponible, es posa a disposició de l'usuari la possibilitat d'utilitzar una funció d'escalat multidimensional anomenat *disparity*.[112]

La figura 6.4 mostra la interfície gràfica dissenyada per al programari *Hyperbolic Space* desenvolupat específicament per a la reducció d'espais multidimensionals a un pla hiperbòlic i la seva representació gràfica.

Figura 6.4: *Interfície gràfica del programari Hyperbolic Space.*



Validació del programari

Una vegada implementat, es valida el funcionament del programari comparant els resultats obtinguts per cadascun dels mètodes amb un conjunt de dades de validació. El conjunt de Fisher[113] s'utilitza normalment com a conjunt de referència en tasques de classificació, i està format per quatre característiques estructurals de flors de tres gèneres diferents d'iris, taula 6.1.

Taula 6.1: *Conjunt d'entrenament de Fisher. Llistat de la llargada (Llar.) i l'amplada (Amp.) de les característiques considerades per als tres gèneres d'iris avaluats.*

Setosa				*Versicolor*				*Virginica*			
Llar. sèpal	**Amp. sèpal**	**Llar. pètal**	**Amp. pètal**	**Llar. sèpal**	**Amp. sèpal**	**Llar. pètal**	**Amp. pètal**	**Llar. sèpal**	**Amp. sèpal**	**Llar. pètal**	**Amp. pètal**
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.1	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.1	1.5	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

Aplicant el programari desenvolupat, l'aplicació del mètode d'optimització PLS permet disposar cadascun dels gèneres d'iris en regions diferents de l'espai $\mathscr{H}^2$ sigui quina sigui la definició de la semblança emprada (de les tres disponibles) i amb independència dels valors inicials, figura 6.5.

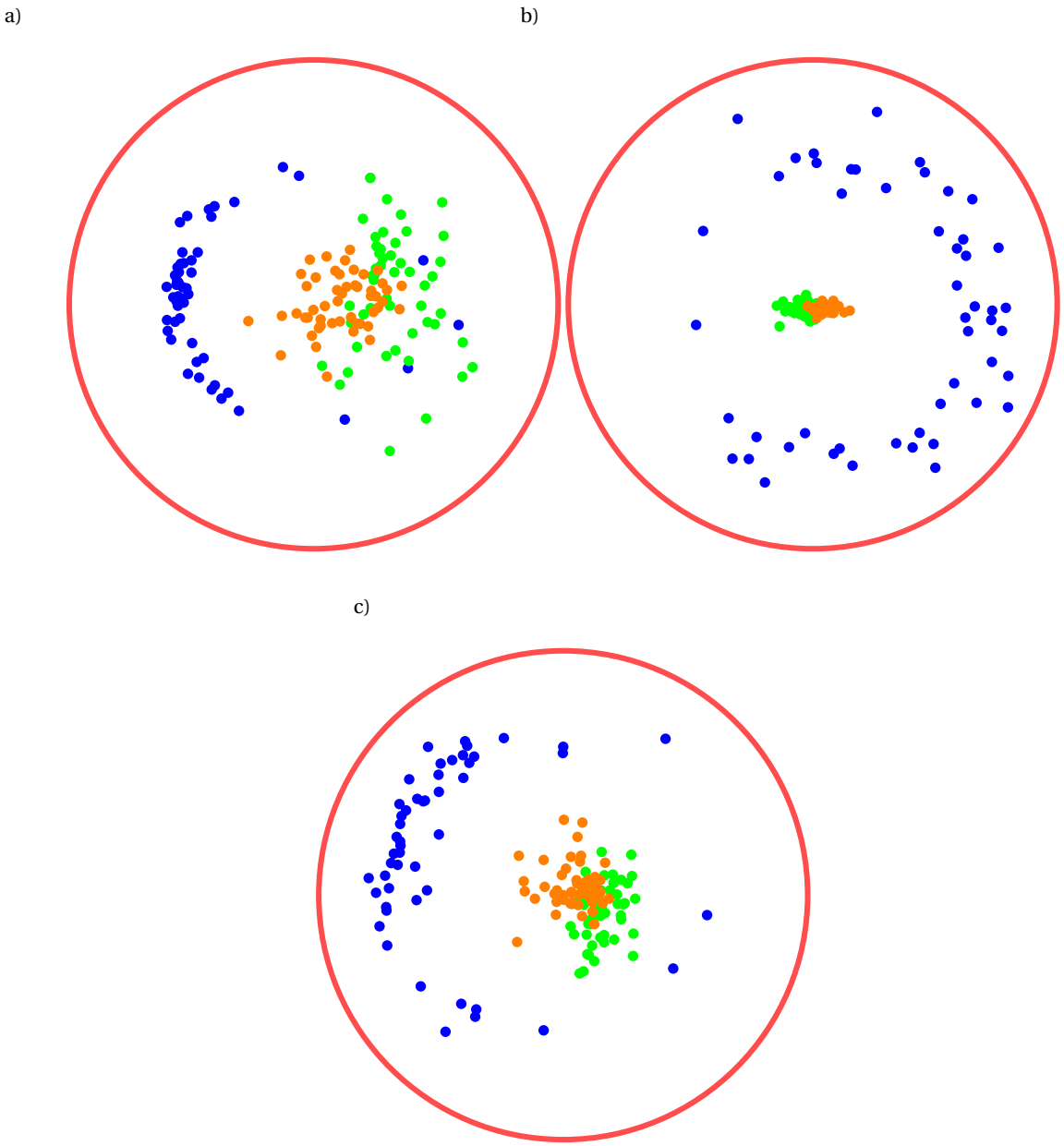


Figura 6.5: *Resultats obtinguts amb PLS en la projecció sobre el pla $\mathscr{H}^2$ del conjunt de Fisher, utilitzant la mètrica euclidiana (a), del cosinus (b) o de Canberra (c). [• iris setosa, • iris virginica, • iris versicolor]*

Per altra banda, la utilització dels algorismes genètics permet aconseguir una separació més homogènia dels punts, figura 6.6.

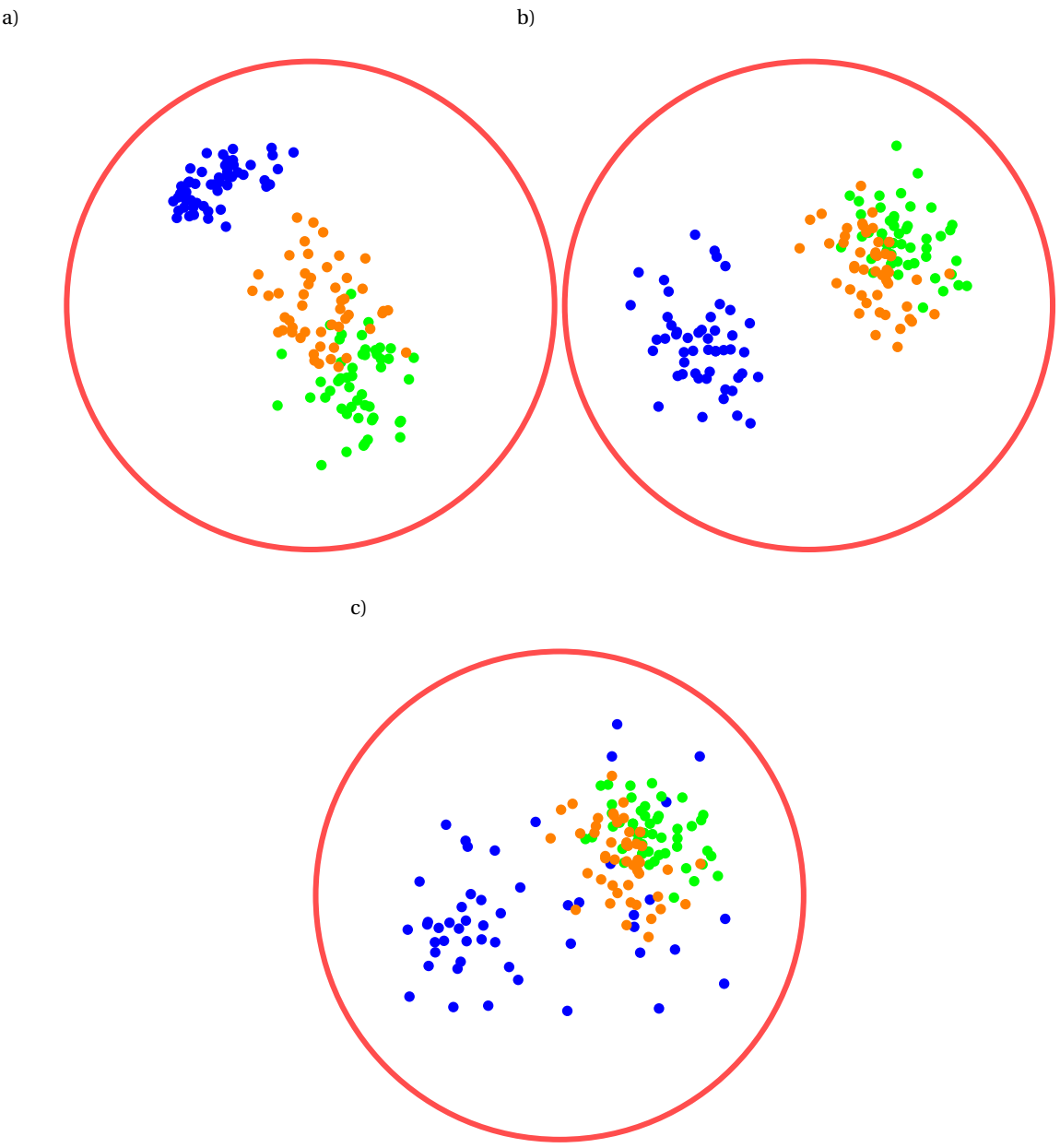


Figura 6.6: *Resultats obtinguts amb GA en la projecció sobre el pla $\mathscr{H}^2$ del conjunt de Fisher, utilitzant la mètrica euclidiana (a), del cosinus (b) o de Canberra (c). [• iris setosa, • iris virginica, • iris versicolor]*

Amb GA s'aconsegueix un major agrupament de les dades i un menor valor de l'error. A més, el patró de distribució sembla ser independent de la mètrica utilitzada, recolzant la idea que els GA són capaços de trobar millors solucions que el mètode PLS. Els resultats obtinguts mostren, per tant, l'agrupació de les dades en funció de les seves característiques, i per aquest motiu, hom pot entendre per què els mètodes que permeten reduir la dimensió de l'espai de descriptors, també esdevenen idonis per realitzar classificacions.

6.5 Properes direccions

Una alternativa, més popular, per reduir la dimensionalitat de l'espai de descriptors són els *Self Organizing Maps*, SOM (pàg. 65). La informació qualitativa que permeten obtenir és equivalent a la projecció sobre el pla $\mathscr{H}^2$, bo i que la seva interpretació no és tan senzilla.

En el grup de recerca s'està actualment treballant en la implementació dels SOM (2D i 3D) en el programari ArIS. Resultats preliminars mostren el gran potencial que poden arribar a tenir aquestes tècniques. Com exemple, i per poder-los comparar amb els resultats anteriors, es realitza un entrenament amb un SOM-2D (20x20, figura 6.7) i un SOM-3D (30x30x30, figura 6.8) sobre el conjunt de Fisher.

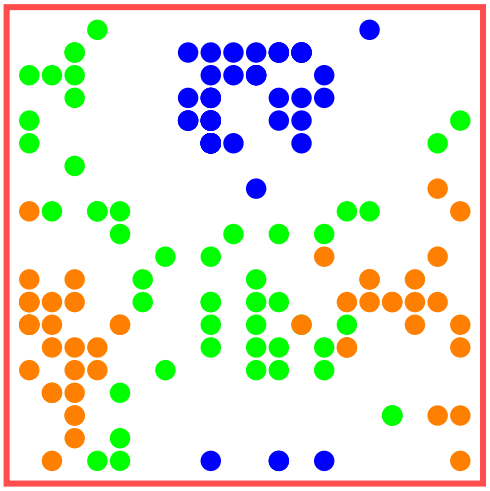


Figura 6.7: *Resultat de la projecció del conjunt de Fisher mitjançant el mètode SOM. [• iris setosa, • iris virginica, • iris versicolor]*

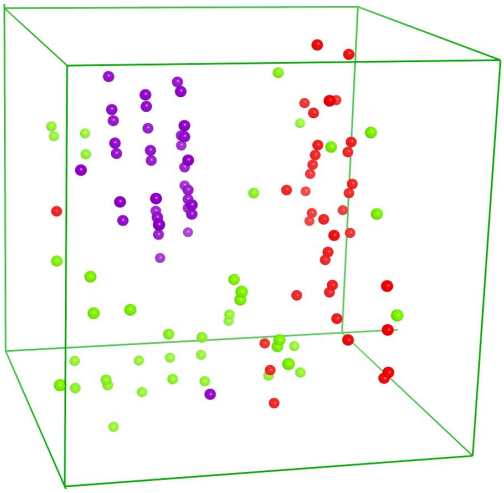


Figura 6.8: *Resultat de la projecció del conjunt de Fisher mitjançant el mètode SOM-3D. La representació gràfica ha estat realitzada mitjançant el programari MOE.*[26] *[• iris setosa, • iris virginica, • iris versicolor]*

Tal com es pot observar, ambdós mètodes són capaços de classificar correctament el conjunt d'entrenament, i a l'igual que la projecció sobre el pla hiperbòlic, poden considerar-se com un mètode de classificació. Després de l'entrenament, hom pot disposar de la matriu de pesos obtinguda en un SOM. Això permet que l'avaluació de noves entrades no hagi de significar tornar a entrenar la xarxa, la qual cosa suposa un gran avantatge sobre el pla $\mathscr{H}^2$.

6.6 Resum

Davant de la dificultat de representar gràficament els espais de descriptors explorats amb la química computacional, el pla $\mathscr{H}^2$ ha resultat un mètode molt eficaç per a la representació gràfica d'espais multidimensionals en dues dimensions. Els resultats obtinguts amb aquesta tècnica són comparables als d'altres tècniques més consolidades, com és el cas dels SOM, però permeten una interpretació més intuïtiva dels resultats.