Universitat de les Illes Balears	ESCOLA POLITÈCNICA SUPERIOR

ESCOLA POLITÈCNICA SUPERIOR
UNIVERSITAT DE LES ILLES BALEARS

PROJECTE DE FINAL DE CARRERA

Estudi:

Enginyeria Informàtica

Títol:

Comparació i classificació de rutes metabòliques

Alumne:
Joan Marc Tudurí Cladera

Director:
Mercè Llabrés Segura

Data: 6 de juny de 2012

Agraïments

A la meva família, la qual m'ha donat el seu suport i ajuda al llarg de la carrera; als companys de la carrera, en especial als "True Sistemes"; als companys de pis, els quals gràcies a la seva excel·lent companyia m'han complicat acabar el project; als Fura; i a na Mercè Llabrés, que a més de ser una gran professora i tutora és sobretot una gran persona.

Resum

El present document considera el cas d'estudi de comparar dues rutes metabòliques. En primer lloc es tracten les dades biològiques i es modelen les rutes metabòliques com a xarxes de Petri. Seguidament es tracten per tenir dades com a models matemàtics i es comparen a partir de coneguts algorismes emprats a bioinformàtica. Finalment, es discuteix a partir de dur a terme un gran nombre de comparacions de rutes metabòliques de distints organismes, quina és la validesa del mètode emprat a l'hora de comparar la relació evolutiva obtinguda amb la relació evolutiva ja estudiada i validada amb altres mètodes.

Índex

Índex de figures

Índex de taules

Introducció

El gran nombre d'avanços tecnològics i científics del segle XX han permès l'evolució de camps com la biotecnologia i alhora la biologia, que s'ha traduït en el descobriment d'una gran quantitat de molècules i altra informació biològica. La seqüenciació del genoma humà al 2003 n'és un exemple que implicà el descobriment d'aproximadament tres mil milions de parells de bases de nucleòtids. Aquests estan organitzats amb el que s'anomenen cromosomes (23 en total amb uns 100 milions de parells de base cada un).

Aquí és on entra en joc la informàtica: davant la gran quantitat d'informació existent es necessita la potència dels computadors per poder tractar i analitzar aquesta de forma ràpida i eficient. Vegem-ho amb un exemple: suposem que com a científic experimentat podem llegir base a base tot el genoma a un segon per base. Amb una jornada laboral de 8 hores i amb 200 hores mensuals, torbaríem un total de 572 anys en analitzar el genoma d'un humà. El cas proposat és una exageració, però sense cap dubte, posa en evidència la necessitat de la informàtica per ajudar a la recerca en biologia.

Què és la bioinformàtica i la biologia computacional? La *bioinformàtica* és l'aplicació de la informàtica i tecnologies relacionades en l'anàlisi, emmagatzemament, manipulació i interpretació de problemes biològics, la base dels quals es troben a les macromolècules com l'ADN, l'RNA i les proteïnes. Inicialment, la tasca d'un bioinformàtic consistia en la creació i manteniment d'una base de dades biològica on emmagatzemar dades com seqüències d'aminoàcids o de nucleòtids. No obstant, el gran avanç del camp de la biotecnologia ha fet créixer nombrosament el volum d'informació biològica, al mateix temps que la informàtica ha evolucionat molt en termes de computació. Per tant, també ha pogut evolucionar una àrea com la bioinformàtica, fent ús d'aquesta potència de càlcul per poder interpretar informació biològica amb eficiència i precisió.

El procés actual d'analitzar i interpretar dades s'anomena *biologia computacional* que juntament amb la bioinformàtica s'encarreguen d'aplicar tècniques informàtiques, matemàtiques i estadístiques per a resoldre problemes

relacionats amb la biologia, com per exemple, la computació genòmica, modelatge molecular i la predicció d'estructures de proteïnes.

Dins el modelatge molecular hi podem incloure l'estudi del metabolisme de les distintes especies o organismes existents a la natura, i en particular, un dels camps que és força interessant és l'estudi de la comparació de xarxes metabòliques. Aquest és un camp multidisciplinar que agafa conceptes de la bioinformàtica i la biologia computacional, tal i com els algorismes d'alineaments de seqüències, amb altres conceptes més matemàtics i formals com són les xarxes de Petri i els algorismes de grafs.

El present document estudia una manera de dur a terme la comparació de rutes metabòliques de distintes espècies. És per això que aquest s'estructura amb les següents parts: en primer lloc es defineixen els conceptes biològics bàsics per que el lector pugui comprendre de que es parla al llarg del document; a continuació es detalla que són les rutes metabòliques, les xarxes de Petri i com podem representar una ruta metabòlica com una xarxa de Petri; seguidament es du a terme una explicació en detall de cada una de les etapes i algorismes que s'empren per dur a terme la comparació de rutes metabòliques; i finalment es discuteixen els resultats d'un conjunt d'experiments de diverses execucions que ens serveixen per validar i analitzar el funcionament del programa.

Capítol 1

Preliminars

1.1 Conceptes biològics

Com que estam fent feina sobre un projecte d'una matèria que conté conceptes de dues disciplines com són la biologia i la informàtica, a continuació definim alguns conceptes biològics per a que els lectors vinguts de la informàtica puguin entendre el document. Tot i això, es parteix de la suposició de que el lector té una base científica general i conceptes bàsics com mol·lècules o àtoms es donen per sabuts.

Proteïna Les proteïnes són mol·lècules formades per cadenes lineals *d'aminoàcids*, les quals es recargolen i formen estructures tridimensionals. Els aminoàcids són mol·lècules orgàniques amb un grup *amino* (`-NH2`), un grup *carboxílic* (`-COOH`, un tipus d'àcid) i una cadena anomenada *radical* d'estructura variable que determina la identitat i propietats dels diferents aminoàcids. Les proteïnes desenvolupen un paper fonamental en la vida i són les mol·lècules més versàtils i abundants dins els organismes. Duen a terme un gran nombre de funcions tals com immunològiques (anticossos), enzimàtiques, hipostàtiques (contribueixen al manteniment del pH), etc.

Enzim Els enzims són un tipus de proteïnes que catalitzen reaccions químiques: un enzim fa que una reacció química la qual té la energia suficient per dur-se a terme però a una velocitat baixa es pugui fer a una velocitat més alta. En aquest tipus de reaccions, els enzims actuen sobre els substrats. Les reaccions que estan controlades per enzims es diuen reaccions enzimàtiques, tot i que en aquest document no farem cap distinció ja que totes les reaccions que tractarem són enzimàtiques.

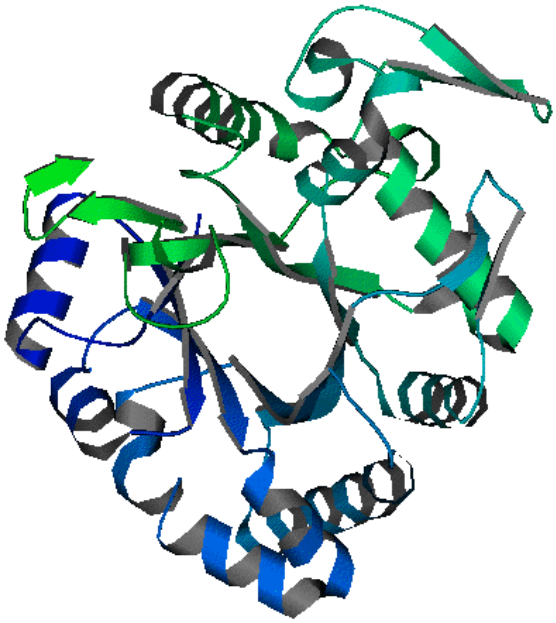

Figura 1.1: *Estructura tridimensional de l'enzim betaglucanasas*

Reacció química Una reacció química és tracta d'un procés on dos o més **substrats** per mitjà d'un factor energètic es transformen en el que s'anomena com a **producte**. Tant els substrats com productes poden ser elements o composts (mol·lècules). Les reaccions químiques solen conduir a canvis que involucren moviment d'electrons formant o rompent enllaços químics.

Metabolisme Es defineix metabolisme com el conjunt de reaccions químiques que tenen lloc dins un organisme per dur a terme funcions vitals. Aquests processos permeten als organismes créixer, reproduir-se, mantenir la seva estructura mol·lecular i respondre als estímuls exteriors. Podem tenir dos tipus de metabolismes. El catabolisme descompon matèria orgànica, com per exemple per extreure energia en la respiració cel·lular. D'altra banda, l'anabolisme utilitza energia per construir components de les cèl·lules com proteïnes i l'ADN.

1.2 Rutes metabòliques

Un cop que ja s'han definit els conceptes biològics necessaris podem explicar que són les rutes metabòliques. Un element important del éssers vius

és el metabolisme, el sistema químic que genera els components essencials (aminoàcids, sucres, acid nucleics...) i la energia necessària per sintetitzar i emprar-los. Anomenam ruta metabòlica a un subsistema del metabolisme d'un organisme la qual desenvolupa una funció específica. Aquestes són una serie de reaccions químiques les quals, a partir d'un conjunt de composts d'entrada (substrats) generen composts de sortida (productes). Sovint aquestes reaccions són catalitzades per un enzim.

Una ruta metabòlica conté un nombre de reaccions encadenades determinat. Es pot donar el cas de que una d'aquestes sigui irreversible, mentre que altres no ho poden ser i el conjunt de reaccions poden succeir-se de manera oposada, depenent de les necessitats de l'organisme. D'aquesta manera, es sol representar com una xarxa de reaccions químiques catalitzades pels enzims on el producte d'una reacció és el substrat de la següent.

L'estudi de les rutes metabòliques requereix combinar informació de distintes fonts: bioquímica, genòmica, anàlisis de xarxes i simulació. Un dels reptes de la computació biològica és representar les rutes metabòliques amb mètodes formals que permetin poder fer anàlisis matemàtics i simulacions estadístiques. Per tal de representar aquesta informació es fa ús de models que permeten un millor coneixement del processos de les rutes tal i com les interaccions entre els components de les rutes i com aquestes contribueixen a la funció i al comportament del sistema sencer. Per aconseguir-ho, s'ha de passar per varies etapes:

1. Traslladar el coneixement teòric i experimental a un model.
2. Validar el model.
3. Derivar del model validat hipòtesis sobre el sistema i verificar-les de manera experimental.
4. Emprar la nova informació adquirida per refinar el model.

A continuació veurem algunes fonts d'informació i bases de dades on s'emmagatzemen models de rutes metabòliques.

1.2.1 Bases de dades de rutes metabòliques

KEGG La base de dades de KEGG (*Kyoto Encyclopedia of Genes and Genomes*) [KEGG, 2012] conté les més conegudes rutes metabòliques de distintes espècies, a més d'integrar informació genòmica, química i informació sistèmica. Al 2012 aquesta conté unes 85.000 rutes, generades a partir de 344 rutes de referencia les qual són dibuixades manualment i actualitzades

sovint. Les rutes són representades per mitjà de mapes amb informació addicional relacionada, tal com proteïnes i gens. Aquests mapes es poden trobar en KGML (*KEGG Markup Language*) basat en tecnologia XML.

SBML i BioModels Una altra base de dades important és la de BioModels, la qual està relacionada amb SBML.org [SBML, 2012]. Permet als biòlegs emmagatzemar i cercar models matemàtics d'interès biològic. Aquests estan vinculats a recursos de interès rellevant com poden ser publicacions o altres bases de dades. Actualment conté 208 models i 27.238 espècies. Aquests models són escrits en llenguatge SBML (*Systems Biology Markup Language*), també basat en tecnologia XML.

MetaCyc Forma part de la base de dades de BioCyc. Descriu més de 1.100 rutes metabòliques de més de 1.500 especies diferents.

Reactome Base de dades on s'emmagatzemen les rutes i les reaccions més importants dels humans.

BioCarta Podem trobar un conjunt de models dinàmics gràfics els quals integren informació protètica de la comunitat científica. Conté unes 80 rutes metabòliques.

1.3 Xarxes de Petri

Una Xarxa de Petri és un formalisme matemàtic, representable gràficament, que es pot fer servir per modelar sistemes de processament d'informació, sistemes distribuïts i en general, concurrents, asíncrons i no deterministes [Peterson, 1981].

1.3.1 Conceptes bàsics de les xarxes de Petri

Una xarxa de Petri (PN, *Petri Net*) finita i marcada és una tupla $N = (P, T, W, M_0)$ on:

- P és un conjunt de llocs (*places*), $P = \{p_1, \ldots, p_n\}$
- T és un conjunt de transicions, $T = \{t_1, \ldots, t_m\}$
- $W : ((P \times T) \cup (T \times P)) \to N$ és la funció de pes, on $W(x, y) = k$, $k > 0$ i la xarxa té un arc que va de x a y amb pes k.

- M_0 és un vector de dimensió $|P|$ d'enters positius que representen el marcatge inicial de la xarxa.

Hem de tenir en compte que $P \cap T = T \cap P = \emptyset$.

Qualsevol element de $P \cup T$ és un node, el qual pot ser un lloc o una transició. Els llocs es representen com un cercle blanc i les transicions com a rectangles negres. Cada lloc de la xarxa conté un nombre enter positiu de marques (representades amb un cercle negre) i es pot veure com un contenidor d'objectes on el nombre de marques són el nombre d'objectes que hi ha dins el contenidor. Una transició es pot veure com un procés que necessita consumir objectes per activar-se i produir nous objectes. Si un arc relaciona un lloc amb una transició significa que les marques d'aquell lloc són les entrades de la transició. Si pel contrari, una transició està connectada amb un lloc, implica que si aquesta s'activa, produirà marques i les inserirà dins el lloc. Els arcs són dirigits i tenen un pes associat (si el pes de l'arc és 1 es pot ometre la indicació) el qual indica el nombre de marques que es consumeixen o es produeixen durant l'activació d'una transició. Cal dir que mai hi pot haver un arc connectat entre dues transicions ni cap arc connectat entre dos llocs.

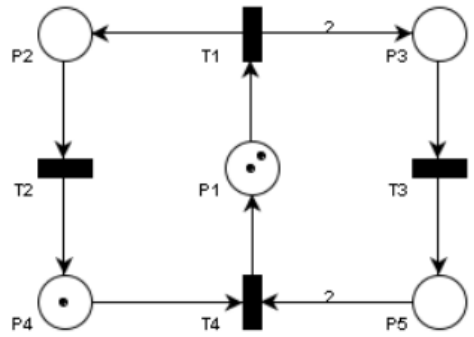

Figura 1.2: *Exemple de xarxa de Petri.*

Un marcatge M és un vector de dimensió $|P|$ d'enters positius que representen la quantitat de marques en cada lloc de la xarxa. Per l'exemple de la figura 1.2 tenim el següent vector M_0:

$$M_o = \begin{pmatrix} 2 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

A més de la representació gràfica de les xarxes de Petri, aquestes poden ser representades amb conceptes matemàtics com són les matrius. Aquest tipus de representació ens serveix per poder veure la xarxa d'una manera formal i posteriorment poder analitzar-la. Un exemple d'aquesta representació és la vista anteriorment per indicar el marcatge d'una xarxa. Anem a veure de quines altres maneres podem representar la xarxa.

Input Matrix La *Input Matrix* (o matriu d'entrades) I és una manera compacta de representar els arcs que connecten els llocs amb les transicions i el pes associat. Té $|P|$ files i $|T|$ columnes. L'element $i_{j,k}$ és el pes de l'arc que va de lloc p_j fins a la transició t_k si l'arc existeix, sinó val 0. Per l'exemple de la figura 1.2 tenim la següent *input matrix*.

$$I = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix}$$

Output Matrix La *Output Matrix* (o matriu de sortida) O ens permet representar els arcs que connecten les transicions amb els llocs i el pes associat. Té $|P|$ files i $|T|$ columnes. L'element $o_{j,k}$ és el pes de l'arc que va de la transició t_k fins al lloc p_j si l'arc existeix, sinó val 0. Per l'exemple de la figura 1.2 tenim la següent *ouput matrix*.

$$O = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Incidence Matrix La *Incidence Matrix* (o matriu d'incidència) C resumeix la informació sobre les matrius *input* i *output*. Té la mateixa dimensió que O i que I i es calcula fent la resta de O i I. Per l'exemple de la figura 1.2 tenim la següent *incidence matrix*.

$$C = O - I = \begin{pmatrix} -1 & 0 & 0 & 1 \\ 1 & -1 & 0 & 0 \\ 2 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -2 \end{pmatrix}$$

Les matrius d'incidència contenen informació tant de les matrius d'entrada com les matrius de sortida però podem perdre cert tipus d'informació. Pot succeir que si tenim situacions tals que $O(j,k) = I(j,k)$ aleshores $C(j,k) = 0$, que pot donar lloc a interpretacions errònies com que el lloc k no està connectat amb la transició j. Aquest cas succeeix quan tenim xarxes amb estructures de bucles (figura 1.3). Les xarxes de Petri que no contenen bucles s'anomenen xarxes de Petri pures, per tant i per aquestes xarxes, podem usar la matriu d'incidència sense perill de tenir ambigüitats.

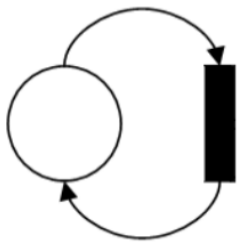

Figura 1.3: *Exemple de bucle.*

1.3.2 Extensions de les de xarxes de Petri

Existeixen diverses extensions del formalisme bàsic de les xarxes descrit anteriorment. A continuació anunciam algunes d'aquestes extensions que certs autors fan servir a l'hora d'emprar xarxes de Petri per fer feina amb rutes metabòliques.

- Xarxes de Petri amb arcs inhibidors. Es tracta d'una xarxa de Petri estàndard amb l'adició d'un tipus d'arc anomenat arc inhibidors els quals permeten activar una transició quan el lloc associat a aquest no té cap marca.

- Xarxes de Petri colorades. Les distintes marques de cada un dels llocs tenen un color associat i els arcs poden ser etiquetats i activar transicions segons els colors.

- Xarxes de Petri temporitzades. S'associa un interval $[i_t, f_t]$ a cada transició t. Quan t s'activa no pot tornar a ser activada fins que han passat i_t unitats de temps i no pot continuar activant-se després de f_t unitats de temps, a menys que t sigui desactivada per un altra transició.

- Xarxes de Petri estocàstiques. En aquest tipus de xarxes, les transicions tenen associades una tassa d'activació que segueix un distribució probabilística.

- Xarxes de Petri continues. En aquestes xarxes el nombre de marques d'un lloc no és discret sinó que pot ser un nombre real no negatiu.

Encara que cada una d'elles té una traducció natural a ruta metabòlica, en aquest projecte només s'usen xarxes de Petri estàndards.

1.4 Modelització de rutes metabòliques com xarxes de Petri

Un cop que s'ha explicat que són les xarxes de Petri convé estudiar per què i com traduir de rutes metabòliques a xarxes de Petri. Malgrat tenim la informació biològica en un format estàndard seguim tenint el problema de que la informació no és l'adequada per dur a terme els esmentats mètodes o algorismes per tractar la informació. Aquesta necessita ser modelada amb el risc de que el model emprat elimini molta informació biològica necessària que és crucial per la investigació. Molts d'aquests models es basen en grafs, amb l'inconvenient de que com més senzill sigui la variant del graf emprada, més pèrdua d'informació biològica tindrem. En contraposició, com més complexe sigui el model emprat, més complex és l'anàlisi d'aquest [Deville et al., 2003].

Una proposta que va a mig camí entre completesa i complexitat són les xarxes de Petri. Com ja hem vist, aquestes tenen una representació gràfica intuïtiva que permet la comprensió del sistema modelat. Les xarxes de Petri tenen la característica de que s'adapten de manera molt natural a la representació de rutes metabòliques ja que hi ha molts de conceptes similars entre sí. Ambdós estan formats per una col·lecció de reaccions les quals consumeixen i produeixen recursos, a més de que la seva representació gràfica es similar. A partir d'això podem aplicar les tècniques desenvolupades per les xarxes de Petri a les rutes metabòliques [Baldan et al., 2010].

1.4.1 Correspondència entre les rutes metabòliques i els elements de les xarxes de Petri

Per tal que una xarxa de Petri representi una ruta metabòlica, el primer pas consisteix en donar una descripció estructural de la xarxa. Aquesta tasca és senzilla ja que la correspondència entre les xarxes de Petri i les rutes metabòliques es bastant natural.

D'una banda, els llocs s'associen amb els compostos moleculars, com els metabòlits, les proteïnes o els enzims. El marcatge de la xarxa té molt a veure, ja que el nombre de marques en cada lloc indica la quantitat de substàncies d'aquest lloc.

D'altra banda, les transicions corresponen a les reaccions químiques. Els llocs situats com a precondició de les transicions representen els substrats o reactants, mentre que els llocs situats a la postcondició d'una transició representen els productes resultants de la reacció. Cal dir que la matriu d'incidència de la xarxa de Petri és idèntica a les matrius estequiomètriques dels sistemes de reaccions químiques.

Ruta metabòlica	Xarxa de Petri
Metabòlits, enzims i compostos	Llocs
Reaccions, interaccions	Transicions
Substrats	Llocs d'entrada
Productes	Llocs de sortida
Coeficients estequiomètrics	Pesos dels arcs
Quantitat de compostos o enzims	Nombre de marques als llocs
Lleis kinètiques de les reaccions	Tasses de dispar de transició

Taula 1.1: *Correspondència entre els elements de les rutes metabòliques amb les xarxes de Petri.*

Quan els enzims són representats com substrats normals, la xarxa de Petri resultant no és pura. Això pot causar algunes dificultats en l'anàlisi de la xarxa que en tenir bucles pot acabar provocant que tenguin comportaments cíclics que no tenen cap interpretació biològica.

Com exemple considerem les dues reaccions de la figura 1.4 les quals formen part de la glicòlisi de KEGG. L'enzim 3.1.3.11 té associada la següent reacció:

$$\text{D-fructose 1,6-bisfosfat} + H_2O = \text{D-fructose 6-fosfat} + \text{fosfat}$$

on els substrats són la D-fructose 1,6–bisfosfat i l'aigua, i els productes són D-fructose 6-fosfat i el fosfat. Cal dir que KEGG empra el signe igual en les

reaccions tant si són irreversibles com si no. La direcció de la reacció s'indica amb fletxes en el diagrama de KEGG.

La figura 1.4 també mostra la reacció inversa, catalitzada per l'enzim 2.7.1.11:

$$\text{ATP} + \text{D-fructose 6-fosfat} = \text{ADP} + \text{D-fructose 1,6-bifosfat}$$

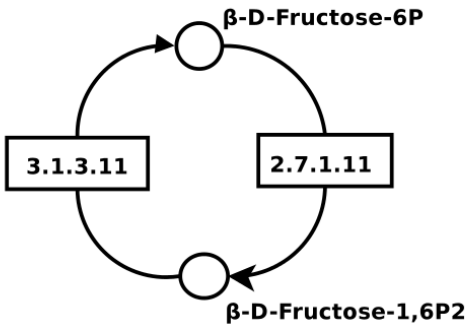

Figura 1.4: *Exemple de ruta metabòlica*

Com ja s'ha dit, podem representar cada compost i cada enzim com un lloc de la xarxa de Petri mentre que les transicions representen reaccions químiques. D'aquesta manera obtenim la xarxa de Petri de la figura 1.5. Es pot observar com el lloc que representa l'enzim està connectat amb la transició amb una doble fletxa, fet que prova que sigui un bucle.

Per tal de simplificar el model de la xarxa i evitar aquests fenòmens es sol modelar evitant la representació explicita dels enzims ometent els corresponents llocs. A més, si observam la figura 1.4 les substancies de l'aigua, els fosfat, l'ADP i l'ATP no es mostren: són molècules ubíqües i s'assumeix que la seva concentració és constant. La figura 1.5 es pot simplificar encara més ometent els llocs d'aquests compostos, resultant en la figura 1.6. Aquestes simplificacions es solen aplicar en general, sobretot per aquelles xarxes molt complexes. Evidentment, al afegir aquesta simplificació, no es representen processos químics relacionats amb els compostos que hem eliminat.

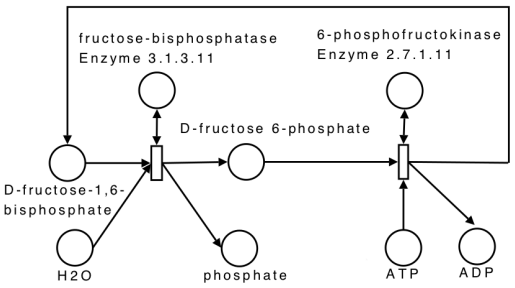

Figura 1.5: *Primera aproximació de xarxa de Petri associada a la ruta metabòlica de la figura 1.4*

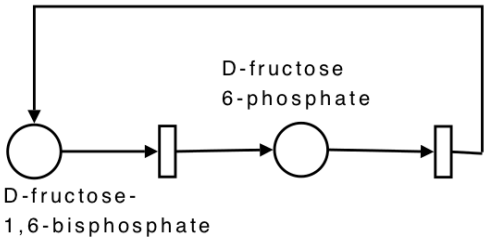

Figura 1.6: *Segona aproximació de xarxa de Petri associada a la ruta metabòlica de la figura 1.4*

1.4.2 Problemes en la representació de rutes metabòliques amb xarxes de Petri

Modelització de propietats espacials En alguns casos, pot ser útil distingir composts d'acord amb la seva localització dins la cèl·lula. Per exemple, dins una cèl·lula, els ATP de dins i de fora dels mitocondris són diferents i la seva concentració relativa és determinada per mitjà d'un procés de transport selectiu. Alguns autors solucionen aquest problema emprant dos llocs diferents per l'ATP [Reddy et al., 1993]. En general, les xarxes de Petri no estan fetes per representar propietats espacials [Hardy and Robillard, 2004]. Conceptes com posicions i distàncies no es solen modelar i la seva modelització no sempre és senzilla. Una altra manera de solucionar aquests problemes és emprant diferents colors en xarxes de Petri colorades [Heiner et al., 2001].

Modelització de metabòlits externs En una ruta metabòlica es poden distingir entre metabòlits interns i externs. Els primers són totalment produïts i consumits a la xarxa mentre que els segons representen fonts o embornals, és a dir, punts de connexió amb altres rutes. Els metabòlits externs essencialment corresponen a la interfície entre la xarxa i l'entorn. Poden ser representats de diferents maneres.

Una primera possibilitat consisteix en simplement no incloure els llocs associats a aquests metabòlits externs. Aquesta aproximació ens pot ser útil a l'hora de fer simulacions.

D'altra banda, si el model explícitament inclou llocs associats als metabòlits externs, llavors estan caracteritzats pel fet de que les transicions o bé totes consumeixen (metabòlits d'entrada) o totes eliminen marques (metabòlits de sortida) en aquests llocs. Una primera solució consisteix en incloure, per cada lloc de la xarxa que es correspon a un metabòlit d'entrada, una transició sense cap precondició, generant així marques en aquest lloc [Heiner and Koch, 2004]. Una transició d'aquestes característiques sempre genera marques, assumint que sempre tenim metabòlits d'entrada disponibles. De manera semblant, podem col·locar una transició sense postcondició per representar els metabòlits de sortida.

Una proposta diferent és emplenar tots els llocs que corresponen als metabòlits d'entrada amb un nombre infinit de marques i permetre als llocs que corresponen als metabòlits de sortida acumular un nombre arbitrari de marques [Zevedei-Oancea and Schuster, 2003].

1.4.3 Què s'ha fet?

A continuació s'enuncien algunes implementacions que s'han fet de diverses rutes metabòliques modelades com xarxes de Petri. En articles com [Reddy et al., 1993] els autors consideren la ruta metabòlica de la *fructosa* i la representen amb una xarxa de Petri. A partir d'ella estudien característiques de les xarxes de Petri tals i com si la xarxa és viva, reversible, l'estructura dels invariants, etc, i a partir d'aquí es discuteixen les seves interpretacions biològiques. També es proposen metodologies algorísmiques per estudiar xarxes complexes de reaccions químiques sense fer ús de la estequiometria. Modelen les rutes com una composició de xarxes de Petri colorades. Posteriorment transformen la xarxa en un dígraf i en una llista de circuits mínims. Això permet la identificació de rutes bioquímiques i l'estudi de flux molecular. Per exemplificar aquesta aproximació fan servir el cicle de *Krebs*, un tipus de ruta metabòlica molt estudiada.

També podem trobar una altra manera que consisteix en emprar eines automàtiques de traducció. El problema d'aquestes és que no fan feina amb

fitxers d'entrada i de sortida de manera estandarditzada. D'una banda, autors com [Shaw et al., 2004] proposen un traductor automàtic on la seva entrada són models SBML i la sortida s'expressa amb un fitxer PNML (*Petri Net Markup Language*). En aquest article es pot veure com s'il·lustra la ruta de la glicòlisi de l'organisme *saccharomyces cerevisiae* obtinguda de la base de dades de BioModels. D'altra banda tenim autors com [Baldan et al.,] que han implementat una eina més completa ja que permet introduir tant fitxers KGML com SBML i la sortida s'expressa amb un fitxer PNML. El problema dels fitxers PNML és que no són un estàndard, per tant, eines com per exemple el PIPE2 [Dingle et al., 2009] no són del tot compatibles amb els fitxers PNML. D'aquesta manera, els autors d'aquesta eina han modificat l'estructura del fitxer PNML perquè pugui ser compatible amb el PIPE2.

Es pot veure que no hi ha cap eina estàndard de traducció de rutes metabòliques a xarxes de Petri. Això passa per dos motius: d'una banda, és difícil modelar qualsevol ruta de manera automàtica, ja que a vegades es necessiten fer les traduccions de manera manual per que tenguin un sentit biològic. D'altra banda, la inexistència d'estàndards i la gran quantitat de bases de dades de rutes metabòliques fa que elaborar una eina que englobi tots els formats sigui una tasca molt complicada.

1.5 Alineament de rutes metabòliques

L'anàlisi de rutes metabòliques pot tenir certa importància a l'hora d'explorar certes propietats d'aquestes tal i com la seva robustesa, els estats d'equilibri, l'estructura modular o bé els motius de la xarxa. Una altra manera d'estudiar les rutes metabòliques és considerant la comparació de múltiples rutes per identificar freqüents subgrafs i el seu alineament. Un alineament no és més que un tipus d'anàlisi comparatiu on a més de proporcionar una puntuació de quan igual poden ser una ruta metabòlica amb altres també ens permet conèixer i identificar quines parts són similars entre distintes rutes metabòliques.

Els alineaments de rutes metabòliques tenen el seu particular interès en l'estudi de malalties i el disseny de fàrmacs per prevenir-les, en la reconstrucció del metabolisme de la seqüenciació de gens per ajudar a la comprensió de les funcions metabòliques, en la reconstrucció filogenètica per descobrir informació evolutiva entre espècies i en l'agrupament d'enzims per identificació d'enzims perduts.

1.5.1 Tipus d'alineaments i propostes d'altres autors

Existeixen distintes propostes per comparar diferents organismes a través de l'estudi del seu metabolisme. Com que una ruta metabòlica és un sistema complexe, cada proposta es basa en una representació simplificada de les rutes metabòliques. D'acord amb les distintes estructures usades per fer l'anàlisi podem trobar els següents tipus d'alineaments o mètodes de comparació de rutes.

Conjunts Una ruta és representada com un conjunt dels seus principals components, els quals poden ser reaccions, enzims i compostos químics. La comparació entre dues rutes consisteix en determinar el nombre d'elements comuns. Una mètrica emprada és l'anomenat *índex de Jacard*. Siguin X i Y dos conjunts que han de ser comparats, l'índex de Jacard és defineix com:

$$J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

A [Forst et al., 2006] la distància entre dues rutes metabòliques, amb topologia idèntica, es defineix com la suma ponderada de les distàncies entre els components de la xarxa. Aquesta distància es calcula per mitjà d'un alineament entre parells de seqüències genòmiques. Les diferències estructurals es calculen per mitjà de l'anàlisi de les matrius d'incidència. Aquestes diferències es tracten col·locant una penalització per *gap* a la distància.

[Liao et al., 2002] proposen comparar xarxes metabòliques de distints organismes per estudiar la seva relació evolutiva. Una xarxa és representada amb un perfil binari mostrant la presència o absència de cada ruta metabòlica. La mesura de similitud entre dos perfils binaris es calcula considerat un patró de coincidència i no coincidència bit a bit, i per mitjà d'un mètode heurístic per tenir en compte les relacions jeràrquiques entre les rutes de la xarxa.

A [Hong et al., 2004] una xarxa metabòlica és dividida en 64 rutes metabòliques. El coeficient de correlació de Pearson s'empra per calcular la similitud entre dos organismes. Finalment es fan servir tècniques de *clustering* jeràrquic dels organismes per obtenir una representació en forma d'arbre filogenètic.

A [Clemente et al., 2005] una ruta metabòlica es representa com un conjunt dels seus compostos, enzims i reaccions, descartant altres metabòlits com l'aigua, ATP, etc. La relació de similitud entre components ha de ser la identitat mentre que per mesurar la similitud entre enzims se fan servir tres tècniques distintes: similitud jeràrquica, similitud per informació de contingut i similitud per ontologia genètica. Es proposa un algorisme heurístic

basat en les distàncies entre els components per així computar la distància entre dues rutes metabòliques.

[Tohsato, 2007] compara xarxes de metabolismes per mitjà d'una cadena de bits que representa la presència o absència d'una reacció. La similitud entre xarxes es fa amb el coeficient de Tanimoto, extensió de l'índex de Jacard, el qual és el nombre de reaccions comunes entre dos perfils de reaccions dividit pel nombre de reaccions que no son comunes.

Seqüències En aquest mètode tenim seqüències de reaccions, descomposant així la ruta en un conjunt de camins que van d'un component inicial a un component final. En aquest cas, per obtenir la puntuació de l'alineament es fan les sumes de les puntuacions parcials obtingudes amb una penalització per *gap* com a mesura de similitud.

[Tohsato et al., 2000] proposa un alineament múltiple basat en similitud de les reaccions. S'obté per mitjà d'un pre-procesament manual el conjunt de camins que mostren similitud entre reaccions d'un conjunt de rutes metabòliques. Cada camí es representat per la seqüència dels corresponents enzims. Finalment s'aplica un alineament múltiple basat en l'algorisme de Needleman i Wunsch [Needleman and Wunsch, 1970] i l'aproximació de Feng i Doolittle [Feng and Doolittle, 1987].

A [Lo et al., 2004] la informació metabòlica de dues espècies diferents A i B són comparades per mitjà de la comparació del conjunt de camins els quals comencen per un punt x i acaben a un punt y. La mesura de similitud es fa de manera semblant a l'índex de Jacard.

[Chen and Hofestadt, 2004] presenten una eina anomenada PathAligner la qual representa les rutes metabòliques com una seqüència d'enzims que descriuen la conversió bioquímica d'un substrat cap a un producte. La comparació entre dues rutes es fa per mitjà per alineament de parells de dues seqüències d'enzims, per mitjà de les puntuacions jeràrquiques entre enzims.

[Li et al., 2008] presenten un *framework* anomenat M-PAS que permet identificar i categoritzar rutes metabòliques conservades de xarxes metabòliques de diferents algorismes. Dues rutes metabòliques són alineades només si les reaccions de cada una d'elles transformen de la mateixa manera els substrats comuns cap productes comuns. La puntuació final s'obté considerant tots els components de la ruta i aquesta pot ser refinada depenent dels interessos biològics de l'usuari.

Grafs La ruta metabòlica es representa com un tipus de graf el qual mostra els components de la ruta metabòlica i la relació que hi ha entre ells. S'empren algorismes complexos que involucren homomorfisme o isomorfisme de

grafs i s'introdueïxen algunes aproximacions heurístiques per reduir el cost computacional.

[Heymans and Singh, 2002] representen les rutes metabòliques com un graf d'enzims el qual és el graf resultant d'eliminar els metabòlits i els substrats de les rutes. La puntuació de similitud s'obté a partir dels enzims, emprant diverses tècniques tal i com la obtenció de la puntuació entre la seqüència d'ADN dels seus gens, la similitud entre les seqüències de proteïnes i la similitud jeràrquica dels enzims.

A [Pinter et al., 2005] s'explica el funcionament de l'eina MetaPathHunter, la qual permet fer una cerca homològica de rutes. Donada una ruta d'entrada i una col·lecció de rutes, l'eina retorna totes les rutes similars de la col·lecció de rutes, donant una puntuació i mesura estadística significant. Les rutes es representen com a grafs on els nodes corresponen a enzims i un arc entre dos nodes és que el producte d'un és el substrat de l'altre. La similitud entre grafs es fa a partir de cerques entre homomorfismes de subarbres. Per tant la topologia queda limitada a un arbre.

[Forst et al., 2006] representen les xarxes metabòliques com un hipergraf dirigit on els nodes són metabòlits i els hiperarcs són els enzims/reaccions. Introdueixen operacions d'unió, intersecció i diferència per formar un àlgebra dels hipergrafs i establir la diferència simètrica com una distància de mesura.

[Wernicke and Rasche, 2007] proposen una eina anomenada MetaPath per cerca homològica: donada una xarxa patró, cerca la similitud de subxarxes en una xarxa amfitriona. Una ruta metabòlica es representada com un graf on els nodes són metabòlits i els arcs són enzims/reaccions. L'algorisme determina si dos subgrafs són homomòrfics si es pot subdividir els seus arcs d'una manera tal que el graf resultant sigui isomorf.

[Ferhat and Kahveci, 2010] proposen SubMAP, un algorisme que identifica alineaments de rutes metabòliques per mitjà de l'alineament d'una a moltes assignacions entre molècules i tenint en compte la similitud topològica entre les xarxes i la similitud homològica de les xarxes, és a dir, la similitud enzimàtica i dels compostos químics.

Xarxes de Petri A la literatura es troben certes referències de modelització i anàlisis de rutes metabòliques amb xarxes de Petri però no ni ha gaire articles dedicats a l'alineament i a la comparació de rutes metabòliques emprant xarxes de Petri. Això és degut a que les xarxes de Petri s'empren principalment per modelitzar sistemes concurrents i dur a terme simulacions amb models matemàtics, de manera semblant com a es fa amb la teoria de cues. Hem vist que les xarxes de Petri tenen una semblança natural amb les rutes metabòliques i les hem emprat de manera semblant a un graf.

Paral·lelament, uns col·laboradors italians proposen comparar dues rutes metabòliques estudiant l'homologia de les reaccions i tenint en compte la similitud dels fluxos elementals, els quals es reflecteixen amb els invariants mínims de la xarxa de Petri.

Capítol 2

Comparació local de rutes metabòliques modelades com xarxes de Petri

Ja hem vist distintes tècniques de modelització de rutes metabòliques a més d'estudiar les diferents maneres de comparar i alinear rutes metabòliques. Del gran conjunt de tècniques estudiades ens centram en fer comparació de xarxes de Petri les quals representaran rutes metabòliques.

Les rutes metabòliques amb les quals feim feina, normalment extretes de la base de dades del KEGG o de BioMod, venen expressades amb un fitxer XML. Aquest fitxer es fàcilment convertible amb una xarxa de Petri per mitjà d'un software anomenat MPath2PN. Aquest ens permet obtenir una representació XML de les xarxes de Petri, anomenat PNML. Aquest tipus de fitxers seran els fitxers d'entrada que feim servir per dur a terme l'alineament.

Sigui P una xarxa de Petri que modela una ruta metabòlica i $\mathcal{R}$ el seu conjunt de camins de reaccions. Considerem el conjunt de tots els camins de reaccions $R_1R_2\cdots R_k$ de P tal que un producte de la reacció R_i és un substrat de la reacció R_{i+1}. Un camí és doncs una seqüència de reaccions.

Per cada xarxa de Petri es calcula el conjunt de camins $\mathcal{R}$ de P a partir d'un algorisme de cerca en profunditat que opera amb els llocs d'entrada per obtenir el camí més llarg: es comparen els camins de reaccions i si un camí està inclòs dins l'altre, només es considera el més llarg.

Seguidament es crea una matriu de similituds entre les reaccions existents a les xarxes P_i i P_j. Per calcular aquesta similitud es considera d'una banda la similitud jeràrquica dels enzims associats a cada reacció i d'altra banda la similitud entre els composts d'entrada i de sortida que intervenen a la reacció per mitjà del software SIMCOMP.

Donades P_i i P_j dues xarxes de Petri, siguin $\mathcal{R}_i$ i $\mathcal{R}_j$ el seus respectius

conjunts de camins de reaccions, la següent passa consisteix en calcular l'alineament de totes les parelles de camins de $\mathcal{R}_i$ amb els camins de $\mathcal{R}_j$ emprant l'algorisme Smith-Waterman a partir de la matriu obtinguda a la passa anterior.

Del conjunt dels alineaments de seqüències de reaccions, per mitjà d'un algorisme de *maximum weighted matching* obtenim la combinació d'alineaments que ens donen puntuació màxima. Aquest algorisme troba la correspondència dels camins de la xarxa N_1 més similars als de la xarxa de N_2. Finalment, definim la puntuació global com la suma de les puntuacions de les seqüències alineades sobre el màxim dels camins de N_1 i N_2.

2.1 Obtenció de les *input* i *output matrix*

Com ja varem dir a la secció 1.3 les *input* i *output matrix* ens expressen una xarxa de Petri de manera compacta, és a dir, ens permeten veure com estan connectats els distints elements d'una xarxa. Aquestes matrius ens serviran per poder conèixer quines transicions estan connectades amb quins llocs i viceversa i així facilitar la feina de trobar els camins de reaccions.

Per obtenir aquestes matrius empram un *parser* de fitxers XML i quan anam reconeixent certs patrons del fitxer XML, anam omplint la *input* o la *output matrix* amb la informació obtinguda. Aquests patrons es corresponen a etiquetes XML:

```
<arc id="id_arc" source="nom_font" target="nom_desti">
```

Segons el nom que trobem al `source` podem saber si l'arc va de transició a lloc o de lloc a transició, per tant, podem anar emplenant una matriu o un altre. Amb l'etiqueta `<value>` continguda dins l'estructura de l'`<arc>` podem saber el pes de l'arc.

2.2 Camins de reaccions de la xarxa

La següent passa del nostre algorisme consisteix en calcular el conjunt de camins de reaccions amb el nombre màxim de reaccions. La construcció de camins es fa a partir de llocs d'entrada de la xarxa, és a dir, els substrats químics que provocaran un conjunt de reaccions en cadena. Quanta més quantitat de camins trobem i més llargs siguin, millor ens definiran una ruta metabòlica i podrem construir alineaments més consistents.

El mètode consisteix en iterar sobre el conjunt de llocs d'entrada de la xarxa. Per cada un d'ells, visitam els seus successors per mitjà d'un esquema

de cerca en profunditat i els anam marcant com a visitats. Aquests successors són les diferents reaccions que processen el substrat d'on partim. Per tant, afegim aquesta reacció al camí. Seguidament comprovam la longitud del camí per posteriorment afegir-ho al conjunt final de camins. D'una banda comprovam si hi ha una altre camí que el contengui, es a dir, que sigui un subconjunt o un subcamí d'ell. En aquest cas no feim res però ja sabem que aquest camí no l'hem d'afegir al conjunt final. Si al contrari, el camí que estem avaluant és un superconjunt (o supercamí) d'un altre, vol dir que aquest el conté i és més llarg, per tant, el substituïm per tal de no deixar camins més curts dins el conjunt de camins resultants. En cas de que no sigui cap de les dues coses, simplement l'afegim al conjunt de camins resultants.

```
camins_reaccions:
    resultat = Buid
    per x en llocs_entrada:
        cami = Buid
        cami_mes_llarg(cami, x)
        visitats = Fals

cami_mes_llarg(cami, lloc):
    visitats[lloc] = Vertader
    per x en i_matrix[lloc]:
        si not visitats[x]:
            visitats[x] = Vertader
            cami.afegir(x)
            per y en o_matrix[x]:
                si no visitats[y]:
                    cami_mes_llarg(cami, y)
            comprova_llongitud(cami)
            cami.elimina(x)

comprova_llongitud(cami):
    existent = Fals
    per i in resultat:
        si cami.essubconjunt(i):
            existent = Vertader
            sortir_bucle
        si cami.es_superconjunt(i):
            resultat.remplaçar(i, cami)
            existent = Vertader
            sortir_bucle
```

```
si no existent:
    resultat.afegir(cami)
```

2.3 Similitud entre reaccions

L'objectiu de la obtenció dels camins de reaccions és tenir una base per poder comparar i alinear distintes rutes metabòliques. Prèviament a aquest pas, necessitam establir una mesura de similitud entre reaccions, és a dir, generar una matriu que ens permeti saber quan similars són dues reaccions entre sí. Aquesta ens permetrà calcular alineaments locals de rutes metabòliques a partir de la similitud entre les reaccions que conformen els camins.

La similitud entre dues reaccions es pot calcular a partir de la similitud entre el enzims que catalitzen les reaccions i comparant els distints composts d'entrada i de sortida. Podem expressar aquesta similitud per mitjà de la següent fórmula:

$$\begin{aligned}\text{SimReacc}(R_i, R_j) = &\ \text{SimEnz}(E_i, E_j) \cdot w_e + \\ &\ \text{SimComp}(I_i, I_j) \cdot w_i + \\ &\ \text{SimComp}(O_i, O_j) \cdot w_j\end{aligned}$$

On SimEnz és la similitud obtinguda entre el conjunt d'enzims E_i i E_j de les reaccions R_i i R_j i SimComp és la similitud dels composts. I_i i I_j fan referència als conjunts de composts d'entrada de les reaccions R_i i R_j mentre que O_i i O_j fan referència als compost de sortida. Els termes w_e, w_i i w_j és la ponderació assignada a cada mesura de similitud. Per defecte feim servir $w_i = w_j = 0.3$ i $w_e = 0.4$.

2.3.1 Similitud entre enzims

Com ja hem comentat cada reacció és catalitzada per un enzim. El nostre objectiu es trobar l'enzim que correspon a una determinada reacció per posteriorment dur a terme un mètode de comparació d'enzims. No obstant, per tal de simplificar el model aquests enzims no es representen a la xarxa de Petri. Per tal d'obtenir els enzims associats a cada reacció l'equip tècnic de KEGG ens proporciona una API que permet trobar aquesta informació per mitjà d'uns *web services*. Per tal de no haver de consultar cada vegada quins són aquests enzims i reduir la latència associada a una petició web hem desat a un fitxer la correspondència entre reaccions i enzims de les reaccions existents a KEGG a mesura que anàvem comparant rutes metabòliques.

El mètode emprat s'anomena similitud per jerarquia EC (*Enzyme Commission)*. Els nombres EC és un esquema de 4 nivells de jerarquia desenvolupat per la *International Union of Biochemistry and Molecular Biology* (IUBMB) i que serveixen per a la representació d'enzims. Es representen amb la notació $a.b.c.d$ on $a, b, c, d \in \mathbb{N}$. Per exemple, l'*arginasa* es numera com 3.5.3.1 el qual indica que l'enzim és un hidrolasa ($3.*.*.*$) i que actua en llaços carbonitrogenats (sub-clase $3.5.*.*$) en les amidines lineals (sub-sub-clase $3.5.3.*$). Els enzims amb una classificació EC són funcionalment homòlegs tot i no tenir la mateixa seqüència d'aminoàcids.

El mètode consisteix en analitzar quants elements comparteixen dos enzims partint d'esquerra a dreta. Per exemple, si tenim l'enzim 4.2.3.2 i l'enzim 4.2.3.7, tendrem una similitud d'un 75%. En canvi, si comparam el primer enzim amb un altre enzim que sigui 4.3.3.2, només tendrem una similitud del 25%, tot i que comparteixen, un 75% dels elements, només es considera el primer ja que s'ha de seguir la jerarquia enzimàtica.

2.3.2 Similitud entre composts

Per dur a terme la comparació d'estructures químiques hem fet servir un software anomenat SIMCOMP(*SIMilar COMPound*) [Hattori et al.,]. Es basa en un algorisme que cerca la subestructura comú de similitud màxima entre dos composts químics. Aquests es representen com un graf bidimensional on el àtoms són vèrtexs i els llaços covalents són arcs. L'algorisme cerca els cliques màxims en el graf d'associació del dos grafs. Un clique és un subgraf d'un graf dirigit tal que per a cada parell de vèrtexs existeix una aresta que els connecta.

La similitud entre dos composts es pot estimar a partir del nombre d'àtoms coincidents en contraposició del nombre total d'àtoms (els no coincidents i els coincidents obtinguts del alineament d'àtoms) [Hattori et al., 2003]. En conseqüència, la puntuació de similitud entre dues estructures químiques es pot formular de manera semblant l'*índex de Jacard* entre dos vectors de bits.

$$T(c_1, c_2) = \frac{B(c_1) \cap B(c_2)}{B(c_1) \cup B(c_2)}$$

On $B(c_1) \cap B(c_2)$ és el nombre d'àtoms en comú que apareixen als vectors que representen els dos composts i $B(c_1) \cup B(c_2)$ és el nombre d'àtoms total que tenen els dos composts.

El problema de comparar composts no es redueix a emprar el SIMCOMP sinó que per cada reacció podem tenir més d'un compost d'entrada i més d'un compost de sortida. Això es tradueix en un problema de seleccionar quin és

la combinació de composts entre dos parells de reaccions que maximitzen la puntuació. Aquest problema es coneix com *maximum matching* i es discuteix a la secció 2.5

A una xarxa hi pot haver una gran quantitat de compostos i en conseqüència molts de compostos a comparar fet que provoca un augment considerable del temps d'execució del programa. De manera semblant al que feim amb els enzims, a mesura que anam executant el programa amb distintes rutes d'entrada, anam desant aquesta informació dins estructures de dades i a un fitxer de text per tal de no fer càlculs redundants constantment.

2.4 Alineament locals de camins de reaccions

L'obtenció de les puntuacions entre distintes reaccions permet afrontar la següent passa sense haver de preocupar-se de cercar la similitud entre reaccions a l'hora d'alinear dues cadenes de reaccions. Amb el que hem obtingut en el procés previ (l'anomenarem *matriu de substitució*) i les cadenes de reaccions podem començar a fer els alineaments.

Donades dues xarxes de Petri P_1 i P_2 i els seus conjunts $\mathcal{R}_1$ i $\mathcal{R}_2$ de camins de reaccions, procedim a alinear tots els elements de $\mathcal{R}_1$ amb $\mathcal{R}_2$, tenint així $|\mathcal{R}_1| \times |\mathcal{R}_2|$ alineaments. El mètode que s'empra per alinear cada parell de camins de reaccions es discuteix a continuació.

2.4.1 Algorisme Smith-Waterman

L'algorisme d'Smith-Waterman [Smith and Waterman, 1981] és un algorisme molt emprat a la bioinformàtica i a la biologia computacional. La comparació de seqüències de proteïnes i seqüències d'ADN es fan sobretot amb aquest algorisme.

Dins els gran nombre de tipus d'alineaments, aquest només permet alinear parells de seqüències i fer alineaments de locals, és a dir, enlloc d'intentar alinear la seqüència sencera (com fan els algorismes globals), l'algorisme va alineant regions similars dins de subseqüències llargues que a nivell global semblen divergents.

Aquest algorisme es va proposar per Temple F. Smith i Michael S. Waterman al 1981. Com altres algorismes de comparació de seqüències, tal com el Needleman-Wunsch, es tracta d'un algorisme que fa servir programació dinàmica. Aquesta és un mètode per solucionar problemes complexes xapant el problema en subproblemes amb la principal avantatge de que molts d'aquests subproblemes són el mateix i per tant, podem solucionar el subproblema una vegada i estalviar temps de computació.

L'algorisme consisteix en calcular una matriu H la qual ens serveix per definir l'alineament local òptim d'acord amb les puntuacions indicades per la matriu de substitució i la puntuació per *gap*. La matriu H és construeix com:

$H(i,0) = 0,\ 0 \leq i \leq |a|$
$H(0,j) = 0,\ 0 \leq j \leq |b|$
La primera fila i la primera columna emplenada de zeros.

si $a_i = b_j$ llavors $s(a_i, b_j) = s(\text{coincidència})$
si $a_i \neq b_j$ llavors $s(a_i, b_j) = s(\text{no-coincidència})$

$$H(i,j) = \max \left\{ \begin{array}{cc} 0 & \\ H(i-1,j-1) + \ s(a_i,b_j) & \text{Coincidència/No-coincidència} \\ H(i-1,j) + \ s(a_i,-) & \text{Eliminació} \\ H(i,j-1) + \ s(-,b_j) & \text{Inserció} \end{array} \right\},$$
$$1 \leq i \leq |a|, 1 \leq j \leq |b|$$

On

- $a, b =$ cadenes de l'alfabet Σ
- $s(x,y)$ ens dona la puntuació entre dos elements $x, y \in \Sigma \cup \{'-'\}$. $'-'$ és la puntuació de *gap*

Vegem a continuació un exemple de com quedaria emplenada la matriu H. Per senzillesa a l'hora d'anomenar els elements de la seqüència hem fet servir una seqüència d'ADN, però podria haver emprat una cadena de reaccions.

```
Seqüència 1 = ACACACTA
Seqüència 2 = AGCACACA
```

Amb la següent informació:

- $s(\text{coincidència}) = 2$
- $s(a,-) = s(-,b) = s(\text{no-coincidència}) = -1$

$$H = \begin{pmatrix} & - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & 2 & 1 & 2 & 1 & 2 & 1 & 0 & 2 \\ G & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ C & 0 & 0 & 3 & 2 & 3 & 2 & 3 & 2 & 1 \\ A & 0 & 2 & 2 & 5 & 4 & 5 & 4 & 3 & 4 \\ C & 0 & 1 & 4 & 4 & 7 & 6 & 7 & 6 & 5 \\ A & 0 & 2 & 3 & 6 & 6 & 9 & 8 & 7 & 8 \\ C & 0 & 1 & 4 & 5 & 8 & 8 & 11 & 10 & 9 \\ A & 0 & 2 & 3 & 6 & 7 & 10 & 10 & 10 & 12 \end{pmatrix}$$

Per tal d'obtenir l'alineament local òptim, es comença amb el valor més alt de matriu (diguem-li (i,j)). Llavors, anem cap enrere cap a les posicions (i-1,j), (i,j-1) o (i-1,j-1) tal que la seva puntuació sigui la més alta. En cas d'empat, prioritza el moviment diagonal. L'algorisme segueix iterant fins que arriba a un cel·la amb valor 0 o a la posició (0,0).

En la matriu d'exemple, el valor més alt correspon a la posició (8,8). L'anada enrere correspon a les posicions (8,8), (7,7), (7,6), (6,5), (5,4), (4,3), (3,2), (2,1), (1,1) i (0,0).

Una vegada que hem acabat, reconstruïm l'alineament de la següent manera: començant des del darrer valor, acabam a la posició (i,j) emprant el camí prèviament calculat. Un bot diagonal implica que és un alineament (bé sigui una coincidència o no). Un bot cap abaix implica que hi ha una eliminació, per tant, posam un *gap*(-). Un bot cap esquerra o dreta implica que hi ha una inserció i també es representa amb un *gap*.

Per l'exemple tenim:

```
Seqüència 1 = A-CACACTA
Seqüència 2 = AGCACAC-A
```

En el nostre cas aplicam el mateix algorisme modificant el paràmetre s de la següent manera:

- $s(R_i, R_j) = SimReac(R_i, R_j)$
- $s(R_i, -) = s(-, R_j) = 0$

Recordem que $SimReac(R_i, R_j)$ és la funció de similitud entre dues reaccions tal i com s'explica en el tema anterior. Cal aclarir que en el nostre escenari, emprant com $s(R_i, -)$ amb penalització per *gap* (-0.2, -0.5,...) els resultats globals obtinguts són els mateixos que sense emprar penalització per *gap*.

2.5 *Matching* d'alineaments màxim i alineament final

L'execució de l'alineament ens proporciona un conjunt $\mathcal{A}$ d'alineaments, el qual es composa de tots el possibles alineaments dels camins de reaccions del conjunt $\mathcal{R}_1$ de la xarxa P_1 amb els de $\mathcal{R}_2$ de la xarxa P_2. A priori podem pensar que el procediment d'alinear dues xarxes metabòliques finalitza en aquest pas. No obstant, necessitam dur a terme una correspondència entre els camins de cada una de les xarxes de tal manera que cada camí estigui alineant amb un i només un camí de l'altra xarxa. L'objectiu es trobar una associació tal i que maximitzi la puntuació global i no seleccionar els alineaments de manera arbitraria.

Per exemple, suposem que tenim una xarxa P_1 amb dues cadenes A i B i una altra xarxa P_2 amb també dues cadenes C, D. Després d'executar l'alineament, la puntuació d'aquestes és la següent:

$$\begin{pmatrix} & C & D \\ A & 1.0 & 0.7 \\ B & 0.7 & 0 \end{pmatrix}$$

Figura 2.1: *Matrius de puntuacions obtingudes al comparar les cadenes A i B amb C i D.*

Si pensam de manera local (i com es faria de manera intuïtiva), seleccionam A amb C i B amb D. Amb aquesta associació veim com la puntuació global és de 1.0. D'altra banda, si intentam maximitzar la puntuació global (i com ho fa l'algorisme que veurem a continuació), seleccionam A amb D i B amb C. Amb aquesta associació obtenim una puntuació de 1.4, per la qual cosa, podem veure que efectivament es una puntuació més alta que la associació anterior.

Per poder dur a terme aquest procés empram un algorisme anomenat *Maximum matching* o correspondència màxima.

2.5.1 Algorisme de *matching* màxim de grafs

Sigui un graf $G = (V, E)$ no dirigit on V és el conjunt de vèrtexs i E el conjunt d'arcs. És defineix un *matching* M com un conjunt d'arcs no adjacents, és a dir, arcs que no comparteixen un vèrtex comú. Es diu que un

vèrtex està emparellat a M si és incident a un arc que pertany al *matching*. En cas contrari es diu que el vèrtex està exposat [Plummer, 2009].

Es diu que un *matching* M és màxim si per un altre matching M' es compleix que $|M| \geq |M'|$. $|M|$ és el *matching* de tamany màxim, és a dir, conté el nombre màxim d'arcs.

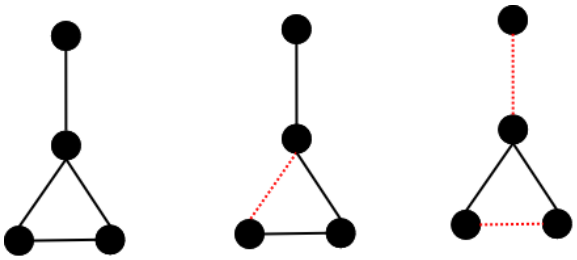

Figura 2.2: *Graf original, un matching del graf i el matching màxim (els arcs amb punts representen els arcs de matching).*

Donat M, es diu que un camí de G és un camí altern si els seus arcs són alternativament a M i no a M o l'inrevés. Un camí augmentador P és un camí altern tal que acaba i comença a dos vèrtexs exposats distints. Es defineix l'augmentació del *matching* a partir d'un camí augmentador P com l'operació de reemplaçar M amb el nou *matching* $M' = M \oplus P = (M \setminus P) \cup (P \setminus M)$

A partir d'aquí es pot aplicar el *Teorema de Berge* per obtenir camins augmentadors: un *matching* M és màxim si i només si no té camins augmentadors. A l'exemple anterior, podem veure com inicialment teníem un *matching* M amb un arc. Al trobar el camí augmentador de M i reemplaçar-ho fins crear M', podem veure com aquest té dos arcs i no té cap camí augmentador, per tant, M' és un *matching* màxim.

Podem tenir una primera aproximació algorísmica:

```
matching_maxim(G, M):
   P = cami_augmentador(G, M)
   si P != Buid:
      M' = augmentacio_matching(M, P)
      retorna matching_maxim(G, M')
   sino:
      retorna M
```

Una primera idea per cercar camins augmentadors consisteix en emprar l'algorisme de cerca de primer en profunditat. Es comença amb un vèrtex exposat com arrel i es va calculant la distància cap l'arrel. Si el node en que

estem està a una distància senar empram un arc de M sinó empram un arc que no sigui de M. Si a qualsevol punt trobam un vèrtex exposat adjacent a un vèrtex de distància parell hem trobat un camí augmentador.

No obstant, aquest algorisme, tot i parèixer que intuïtivament funciona bé, pot provocar situacions en que no trobem solució. Suposem que tenim el següent graf (figura 2.3, si començam el camí amb la seqüència 6, 5, 4, 3, 2, 1 trobam un camí augmentador. D'altra banda, si començam el camí per 6, 2, 3, 4, 5 no trobam el camí augmentador. Aquest problema és provocat per la presència de cicles de longitud senar.

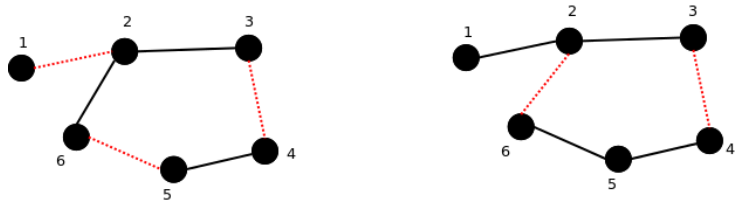

Figura 2.3: *Graf amb el camí augmentador trobat i amb el no trobat.*

Algorisme de *matching* màxim per grafs en general Es defineix *blossom* (o flor) com un cicle de longitud senar ($2k + 1$ arcs) amb k arcs d'un *matching* M.

El fonament d'aquest algorisme és un lema el qual diu que: sigui M un *matching* d'un graf G i B un *blossom*, assumim que B és vèrtex-disjunt de la resta de M i considerem el graf G' com el graf obtingut després de contreure B a un simple vèrtex. Aleshores, el *matching* M' de G' és màxim si i només si M és màxim a G. La operació de contreure consisteix en passar d'un arc a un vèrtex (o passar de dos vèrtex a un de sol).

```
matching_maxim(G):
   M = selecciona_matching(G)
   fer
      etiquetar_vertexs
      si existeix(blossom):
         contreu_blossoms
         continuar
      sino
         P = trobar_camins_augmentadors_disjunts
      M = augmentacio_matching(M, P)
   mentre existeix(blossom) o existeix(cami_augmentador)
   expandir_tots_blossoms
```

Es defineix un bosc com una estructura de dades composta per arbres units de manera disjunta. D'aquesta manera, l'algorisme processa tots els arcs que són fora del bosc i causa la expansió del bosc. Seguidament, (figura 2.4), detecta un *blossom* i contreu el graf.

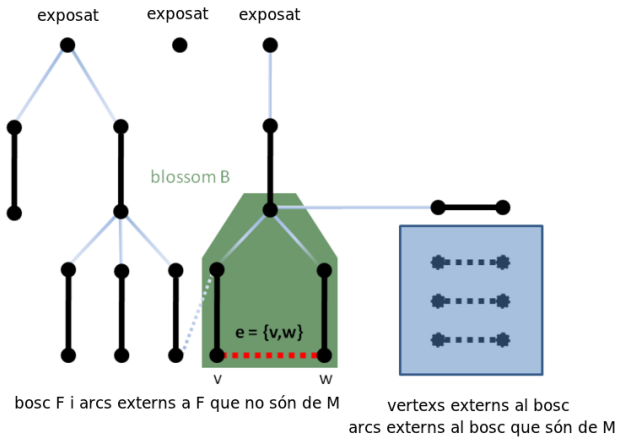

Figura 2.4: *Contracció dels blossoms.*

Finalment localitza un camí augmentador P' al graf contret (figura 2.5) i expandeix fins a obtenir el graf original (figura 2.6). Podem veure la importància de contreure els *blossoms* a l'algorisme ja que aquest no pot trobar P al graf original directament perquè només es consideren els arcs que són fora del bosc entre els vèrtexs a distància parell des de l'arrel.

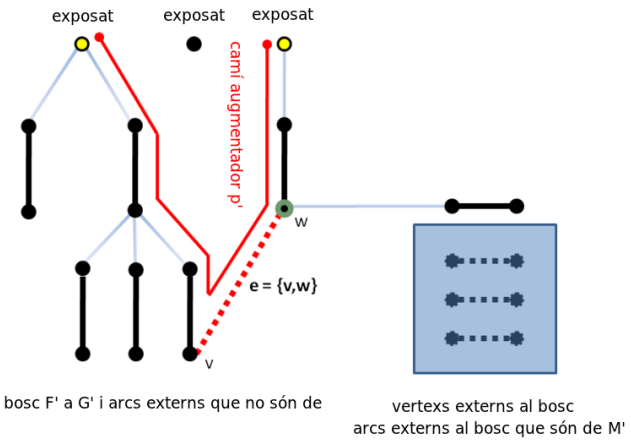

Figura 2.5: *Detecció del camí augmentador.*

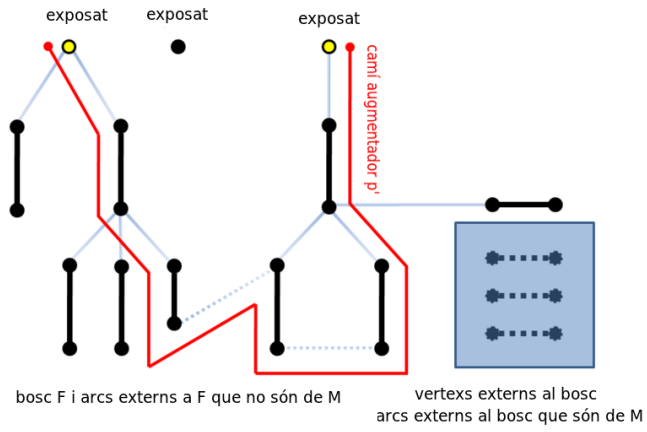

Figura 2.6: *Expansió del blossoms.*

Matching ponderat El problema del *matching* es pot generalitzar a un graf on els arcs tenen pesos i consultar quins dels possibles *matching* M produeixen el *matching* amb un pes total màxim. Aquest problema es pot solucionar amb un algorisme combinatori que empra l'algorisme vist anteriorment com a subrutina [Plummer, 2009].

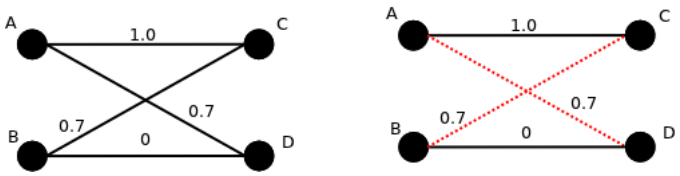

Figura 2.7: *Representació amb un graf de l'exemple anunciat per mitjà de la matriu de la figura 2.1.*

Amb això tenim assignat un *matching* del conjunt de camins $\mathcal{R}_1$ al conjunt de camins $\mathcal{R}_2$, i per tant podem definir la puntuació final entre les dues xarxes.

2.5.2 Resultats finals

La puntuació final que podem obtenir després d'haver fet el *matching* dels alineaments es pot calcular com:

$$S(P_1, P_2) = \frac{\sum \sigma(\mathcal{R}_1^i, \mathcal{R}_2^j)}{max\{|\mathcal{R}_1|, |\mathcal{R}_2|\}}$$

On $\sigma(\mathcal{R}_1^i, \mathcal{R}_2^j)$ són les puntuacions de correspondència dels alineaments obtingudes pel *matching*.

Així obtenim l'alineament del conjunt de camins de reaccions de cada una de les xarxes que hem comparat.

Capítol 3

Experiments, resultats i discussió

Un cop que s'ha explicat el funcionament del software que du a terme la comparació de rutes metabòliques cal mostrar i discutir els resultats obtinguts a partir de l'execució d'un conjunt d'experiments basats en distints organismes.

En primer lloc es mostra un exemple del procés sencer de comparació de dues rutes metabòliques. Seguidament es duen a terme uns experiments meś complicats amb un volum de dades més gran i es discuteixen els resultats obtinguts. Finalment acabam analitzant alguns paràmetres estadístics obtinguts després d'executar l'algorisme repetides vegades.

3.1 Procés de comparació de dues xarxes

Per exemplificar el procés de comparació de xarxes metabòliques farem servir dues xarxes de la base de dades de KEGG, en aquest cas, compararem la de l'ésser humà amb la mateixa ruta de la vaca.

3.1.1 Xarxes obtingudes de KEGG

Per fer-se una idea general de que són les rutes metabòliques, cal mostrar com es representen aquestes a la pagina web del KEGG. A la figura 3.1 es mostra la ruta de la glicòlisi on els rectangles representen els enzims/reaccions mentre que els cercles representen els substrats i els productes.

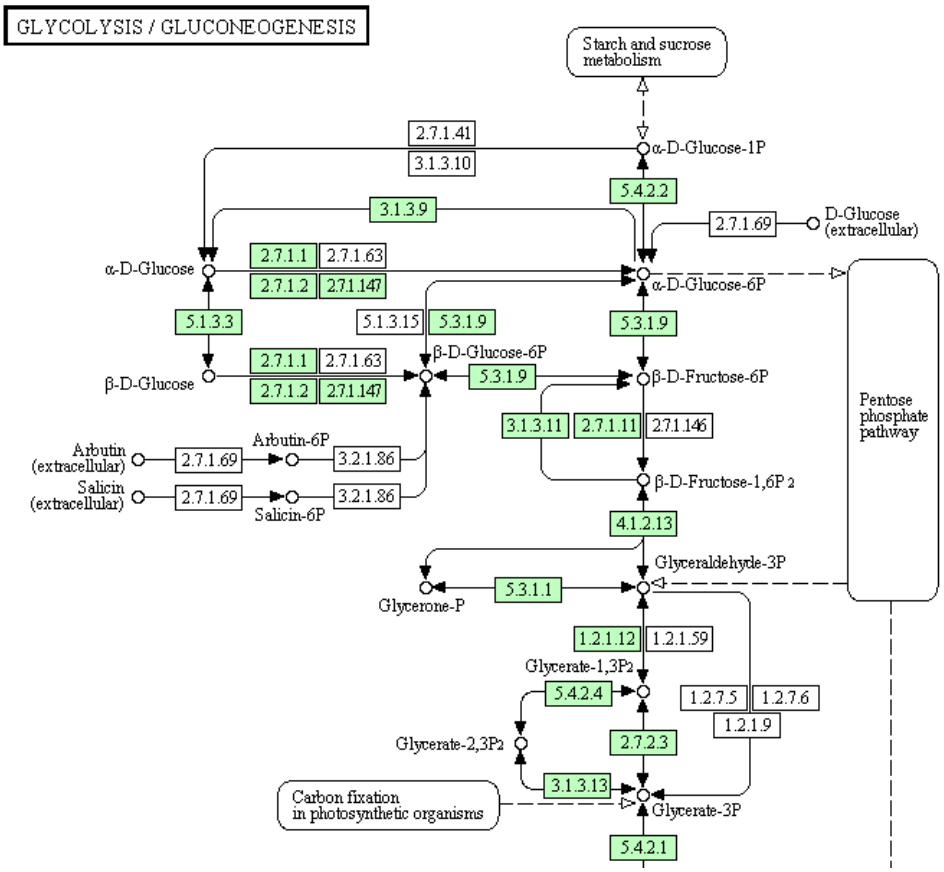

Figura 3.1: *Representació gràfica de la ruta de la glicòlisi de la base de dades de KEGG.*

3.1.2 Representació en forma de xarxa de Petri

Un cop que tenim les xarxes en format XML, podem transformar-les a xarxes de Petri, és a dir, al format PNML. Aquestes ens permetran posteriorment poder obtenir les matrius *input* i *output* per poder obtenir els camins de reaccions. A la figura 3.2 podem veure una part de la representació gràfica de la xarxa en format PNML interpretada pel visualitzador de xarxes de Petri PIPE.

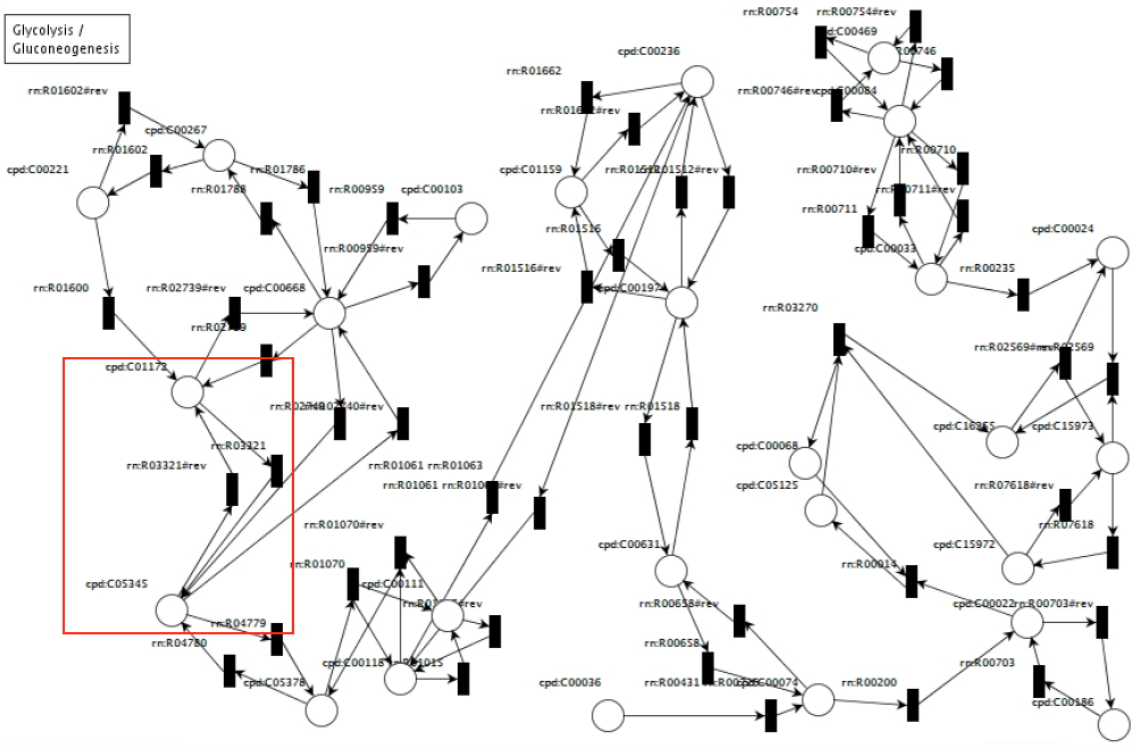

Figura 3.2: *Representació en xarxa de Petri de la ruta de la glicòlisi.*

3.1.3 Comparació i puntuació final

Un cop que hem processat tota la informació en brut i tenim les estructures de dades necessàries, procedim a realitzar la comparació. D'aquest obtenim una matriu amb tots els alineaments de conjunts de camins i la seva puntuació obtinguda tal i com ja s'ha explicat. Per exemple, podem veure com la vaca i els humans conservam una seqüència comuna de reaccions:

```
R00431 R00726 R00200 R00703#rev R00703
R00431 R00726 R00200 R00703#rev R00703
puntuació 1.0
```

La puntuació en aquest cas és 1, ja que les seqüències són idèntiques. L'algorisme també ens genera el següent alineament:

```
R00431 R00726 R00200 R01196#rev  -        -      R01196
R00431 R00726 R00200 R00014    R03270 R02569#rev R02569
puntuació 0.5234685
```

Aquí podem veure com aquests dos camins tans sols tenen un similitud del 0.52. Aquest fet és provoca ja que un camí és més curt que l'altre, per tant, s'introdueixen gaps i fa que la puntuació entre ambdós camins sigui més petita.

Finalment, després d'aplicar els algorismes de *Matching* d'alineaments màxim obtenim tots els parells d'alineaments tal i que la puntuació és màxima. En aquest cas la vaca i l'ésser humà per aquesta ruta metabòlica es pareixen en un 47.98%.

3.2 Anàlisis del recobriment de les xarxes

Un anàlisi interessant a observar és quina és la relació entre la longitud mitja dels camins obtinguts a partir d'una xarxa, amb la quantitat de reaccions distintes que té aquesta i de que ens serveix amb relació a la mesura de similitud obtinguda al comparar xarxes.

Per dur a terme un experiment de tals característiques s'han agafat 8 organismes i s'ha fet una comparació de tots amb tots emprant la ruta metabòlica de la glicòlisi. A continuació es mostren dues taules on es pot veure, d'una banda, el nombre de reaccions distintes que intervenen en la comparació de les rutes metabòliques i d'altra banda una taula amb el nombre de camins i la longitud mitja d'aquests en cada ruta (taules 3.2 i 3.3).

codi	organisme	classificació
APE	A.pernix	Archaea
ECO	E.coli	Bacteri
HSA	H.sapiens	Eukaryota
MMU	M.musculus	Eukaryota
SCE	S.cerevisiae	Eukaryota
STY	S.enterica	Bacteri
DME	D.melanogaster	Eukaryota
HPY	H.pylori	Bacteri
PAI	P.aerophilum	Archaea
SPY	S.pyogenes	Bacteri
TKO	T.kodakaraensis	Archaea

Taula 3.1: *Organismes emprats per estudiar el recobriment de les xarxes.*

	ape	eco	hsa	mmu	sce	sty	dme	hpy	pai	spy	tko
ape	-	45	53	53	44	45	49	40	38	42	40
eco	-	-	57	57	48	53	53	47	48	46	49
hsa	-	-	-	50	52	57	50	52	53	56	55
mmu	-	-	-	-	52	57	50	52	53	56	55
sce	-	-	-	-	-	48	48	46	53	50	48
sty	-	-	-	-	-	-	53	47	48	46	49
dme	-	-	-	-	-	-	-	48	49	52	51
hpy	-	-	-	-	-	-	-	-	40	46	35
pai	-	-	-	-	-	-	-	-	-	44	40
spy	-	-	-	-	-	-	-	-	-	-	45
tko	-	-	-	-	-	-	-	-		-	-

Taula 3.2: *Nombre de reaccions distintes que intervenen en la comparació de les rutes.*

	ape	eco	hsa	mmu	sce	sty	dme	hpy	pai	spy	tko
Nombre de camins	24	70	21	21	18	70	18	11	8	70	10
Nombre de reaccions	34	43	50	50	42	43	46	28	33	38	27
Longitud mitja	5	7	7	7	7	7	8	3	3	7	5

Taula 3.3: *Nombre de camins, nombre de reaccions i longitud mitja dels camins de cada organisme.*

De la primera taula podem observar que la mitja de reaccions que intervenen en la comparació de la ruta metabòlica de la glicòlisi és de 48 reaccions. A partir de la segona taula podem extreure que la longitud mitja dels camins d'una xarxa és de 6, tenint un mínim de camins de longitud 3 a algunes xarxes i un màxim de camins de longitud 8, essent els camins de longitud 5 i 7 els més abundants.

Els autors de l'article de SubMAP [Ferhat and Kahveci, 2010] en el seu experiment arriben a la conclusió de que per comparar rutes metabòliques tan sols es necessari obtenir subestructures de màxim 4 reaccions. D'altra banda, nosaltres podem obtenir cadenes d'una mitja de 6 reaccions i de com a molt 8 reaccions. Per tant, podem alinear cadenes més llargues i observar en millor detall els camins comuns entre les dues xarxes. Recordem que el concepte de subestructura per ells és similar als camins que definim nosaltres.

3.2.1 Relació entre nombre de camins i puntuació global

La següent pregunta a discutir és si la longitud dels camins provoca que els alineaments siguin penalitzats.

	ape	eco	hsa	mmu	sce	sty	dme	hpy	pai	spy	tko
ape	1	0.31	0.66	0.66	0.59	0.31	0.60	0.39	0.33	0.23	0.37
eco	-	1	0.26	0.26	0.21	1	0.24	0.13	0.10	0.78	0.12
hsa	-	-	1	1	0.74	0.26	0.85	0.19	0.14	0.16	0.44
mmu	-	-	-	1	0.74	0.26	0.85	0.19	0.14	0.16	0.44
sce	-	-	-	-	1	0.21	0.92	0.20	0.15	0.13	0.47
sty	-	-	-	-	-	1	0.24	0.13	0.1	0.78	0.12
dme	-	-	-	-	-	-	1	0.21	0.27	0.14	0.48
hpy	-	-	-	-	-	-	-	1	0.72	0.15	0.29
pai	-	-	-	-	-	-	-	-	1	0.11	0.52
spy	-	-	-	-	-	-	-	-	-	1	0.13
tko	-	-	-	-	-	-	-	-		-	1

Taula 3.4: *Puntuació obtinguda de comparar els organismes tots amb tots.*

Observem el cas de la puntuació obtinguda a partir de comparar els organismes *ape*(*Aeropyrum pernix*) amb *hsa* (humà). La similitud entre ambdós organismes és d'un 66.5%. Anem a veure alguns camins alineats.

```
341 658#r 1518 1512 1061 1063#r 1015 1070#r 4780 2740#r 2739 2739#r
431 658#r 1518 1512 1061 1063#r 1015 1070#r 4780 2740#r 2739 2739#r
puntuació 1.0
```

En aquest cas veim que és un clar exemple de que el fet de tenir un nombre més llarg de camins dóna puntuacions altes. Anem a veure un altre exemple:

```
235 2569 2569#rev 7618#rev   -
431 0726 200      703#rev   703
puntuació 0.337
```

En aquest altre cas, el que podem veure és que amb dos camins curts la puntuació resultant és molt baixa. Una possible explicació a això és que com més llarg siguin dos camins, si hi ha reaccions que no coincideixen afegirà una penalització a la puntuació total que no serà molt significativa ja que al

calcular la puntuació final feim la mitja de les puntuacions obtingudes dividint pel nombre màxim de reaccions entre el dos camins. Per tant, la mitja de 6 reaccions per camí sembla ser un nombre adequat per alinear camins ja que dificulta que hi hagi penalitzacions com les comentades anteriorment.

També penalitza el resultat global de la comparació de xarxes alinear camins llargs amb camins petits, i sobretot alinear xarxes amb conjunts de camins en cardinal molt diferents, també afecta a la puntuació.

3.2.2 Relació entre nombre de reaccions i puntuació global

Un altre estudi interessant a realitzar és comprovar quina és la relació entre el nombre mig de reaccions entre dos organismes i la puntuació que se n'obté de comparar-los. Per fer tal experiment, s'ha fet una dispersió de punts amb les dades obtingudes a la taula 3.2 i les dades de la taula 3.4. El resultat es pot observar a continuació:

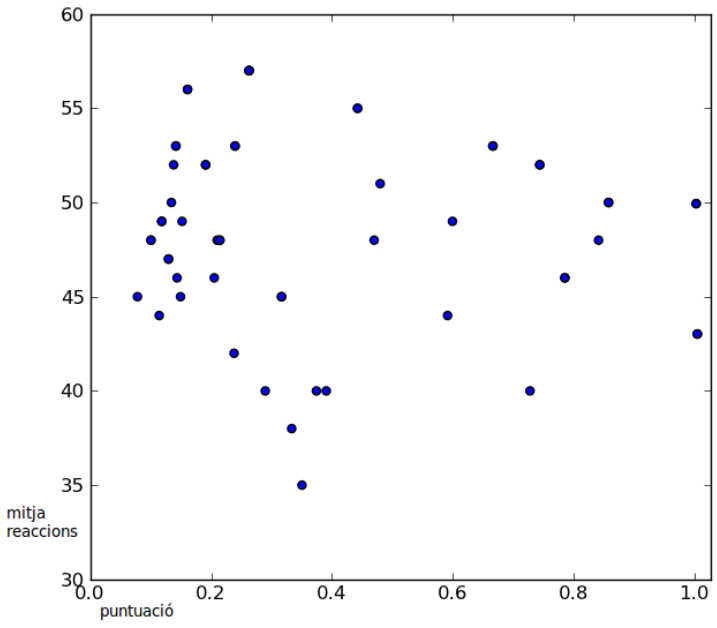

Figura 3.3: *Gràfic de la dispersió de punts obtinguda al comparar la mitja de reaccions amb la puntuació.*

Com es pot veure, la dispersió no segueix cap patró, sinó que més bé la

distribució d'aquests és arbitrari. De fet, si calculam el coeficient de correlació de les dues variables obtenim un valor de -0.081, el qual ens diu que no hi ha cap tipus de correlació entre la quantitat de reaccions entre dues xarxes i la puntuació.

3.3 Experiments i discussió

Un cop que ja s'ha explicat el funcionament passa a passa de l'eina de comparació de xarxes cal posar a prova aquesta amb un conjunt d'organismes més gran. En aquest cas, agafam el mateix conjunt d'organismes de l'article [Clemente et al., 2005] per tal de veure les similituds i diferències amb el nostre algorisme i l'algorisme que han fet servir els autors de l'article. De nou hem fet servir com a xarxa metabòlica de referència la *glicòlisi* degut a que és una xarxa molt estudiada dins la comunitat biològica. Cal aclarir que tot i que l'article citat faci la prova amb 73 organismes nosaltres tan sols la feim amb 53, ja que hi ha organismes que no hem trobat a la base de dades de KEGG o que la seva representació en xarxa de Petri resultant és un graf disconnex no apte per fer comparacions amb altres xarxes.

A la taula 3.5 podem veure el codi de l'organisme i el seu nom científic.

Quan el conjunt d'organismes que volem comparar té unes dimensions com l'exemple estudiat, no podem representar tot el conjunt amb una matriu. Una manera de veure la relació de similitud dels distints organismes es per mitjà d'un dendrograma creat a partir d'un algorisme de *clustering* en concret. Això també ens permet poder inferir quines possibles relacions evolutives hi ha entre els distints organismes a partir de la comparació de les seves rutes metabòliques.

Podem veure l'arbre filogenètic resultant a la figura 3.4 on hi distingim 6 grups diferents d'organismes. En el grup 1 podem veure organismes de tipus *eucariota*, és a dir, el grup on pertanyen les plantes, els animals i el fongs. En aquest grup tenim sobretot animals com els éssers humans, les rates o les mosques. En el grup 2 i 3 ens trobam sobretot bacteris, els qual pertanyen al regne dels organismes *procariotes*. En aquest grup podem trobar bacteris com el de la tuberculosi, la lepra o la *Escherichia coli* un tipus de bacteri molt estudiat i que sol provocar diarrea i altres malalties gastrointestinals. En el grup 5 també podem veure altres bacteris com tots els bacteris relacionats amb la clamídia. Finalment, en els grups 4 i 6 podem veure uns altres tipus d'organismes anomenats *archaea* de tipus procariota i semblants als bacteris.

Comparant amb l'article citat anteriorment, podem veure en primer lloc que el nombre de grups que hem obtingut es bastant menor que al del experiment. Si ens fixam amb l'arbre central de la figura 2 (valor $\alpha = 50\%$)

veim com hi ha grups que coincideixen amb els nostres, com per exemple el grup 5 compost per els organismes *cpa, cpn, cpj, cmu, ctr* i els subgrups *ecs, eco, ece, ecj* i *lmo, lin, cac* dins el grup 3. D'altra banda, veim com el grup dels eucariotes (grup 1) està present a l'arbre canònic del NCBI (figura 1, a l'esquerra). A partir d'aquestes similituds, podem considerar que el nostre experiment valida la comparació de rutes metabòliques definida i implementada, i que la classificació obtinguda és bona.

codi	organisme
ATH	A.thaliana
CEL	C.elegans
DME	D.melanogaster
HSA	H.sapiens
MMU	M.musculus
RNO	R.norvegicus
SCE	S.cerevisiae
SPO	S.pombe
ANA	Anabaena
ATC	A.tumefaciens C
ATU	A.tumefaciens
BHA	B.halodurans
BME	B.melitensis
BSU	B.subtilis
CAC	C.acetobutylicum
CJE	C.jejuni
CMU	C.muridarum
CPA	C.pneumoniae AR39
CPJ	C.pneumoniae J138
CPN	C.pneumoniae
CTR	C.trachomatis
DRA	D.radiodurans
ECE	E.coli O157
ECJ	E.coli J
ECO	E.coli
ECS	E.coli O157J
FNU	F.nucleatum
HIN	H.influenzae
HPJ	H.pylori J99
HPY	H.pylori
LIN	L.innocua
LLA	L.lactis

codi	organisme
LMO	L.monocytogenes
MLE	M.leprae
MLO	M.loti
MTC	M.tuberculosis CDC1551
MTU	M.tuberculosis
PAE	P.aeruginosa
PMU	P.multocida
RPR	R.prowazekii
RSO	R.solanacearum
SAU	S.aureus N315
SAV	S.aureus Mu50
SCO	S.coelicolor
SME	S.meliloti
SPN	S.pneumoniae
SYN	Synechocystis
TMA	T.maritima
TTE	T.tengcongensis
VCH	V.cholerae
XCC	X.campestris
YPE	Y.pestis
AFU	A.fulgidus
APE	A.pernix
MAC	M.acetivorans
MJA	M.jannaschii
MMA	M.mazei
MTH	M.thermoautotrophicum
PAB	P.abyssi
PAI	P.aerophilum
PFU	P.furiosus
SSO	S.solfataricus
MLO	M.loti

Taula 3.5: *Taula del organismes emprats per dur a terme l'experiment amb la ruta de la glicòlisi.*

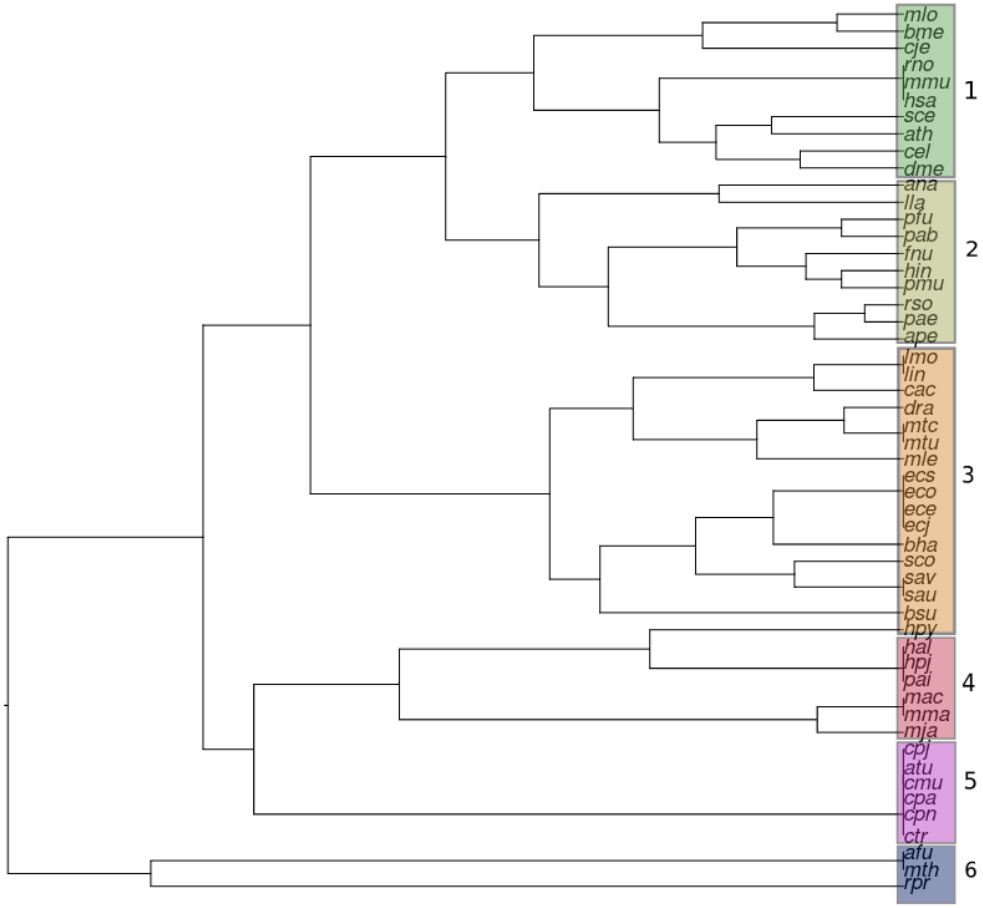

Figura 3.4: *Arbre filogenètic resultant donant com entrada la matriu de resultats de comparar els 53 organismes.*

Conclusions

Un cop finalitzat aquest projecte ens feim a la idea de l'enorme utilitat que té la bioinformàtica per atacar problemes biològics. En el cas de la comparació de sistemes biològics, si bé no és un tema molt pràctic a diferència de l'alineament de proteïnes, es tracta d'un tema suficientment complexe per que sigui interessant desenvolupar noves solucions de les ja existents a la literatura.

Pel que fa al nostre cas hem emprat les xarxes de Petri com a mètode per resoldre un problema biològic que no té solució trivial: comparar rutes metabòliques. El fet d'emprar aquest mètode no provoca que els resultats obtinguts puguin ser millors que emprant altres eines de modelatge de sistemes biològics, però si que és cert que representar rutes metabòliques com xarxes de Petri permet dur a terme una representació més estructurada que no pas emprant grafs normals. A més, modelar amb xarxes de Petri permet que en el futur es puguin emprar altres eines analítiques per poder comparar i alinear rutes metabòliques.

Podem concloure dient que el resultats obtinguts de comparar certs organismes són els resultats que es poden esperar si feim una comparació d'aquests organismes amb altres mètodes com l'alineament de seqüències d'ADN. No obstant, a vegades hem obtingut resultats no desitjats que no és tan la inexactitud del mètode emprat sinó que també és per mors del procés de modelatge o per imperfeccions durant el procés d'obtenció de la pròpia ruta metabòlica. Sigui com sigui el resultat, el fet és que es tracta d'una àrea molt interessant i que els nous avanços en la investigació d'aquesta matèria permeten que en un futur les dades obtingudes siguin més fiables i alinear rutes metabòliques sigui un camp de gran interès dins la biologia i dins la biologia computacional.

Apèndix A

Publicacions

- Mercè Llabrés Segura i Joan Marc Tudurí Cladera. Local piecewise alignment of metabolic pathways. *The Third International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies, BIOTECHNO 2011, May 22-27, 2011 - Venice/Mestre, Italy.*

Bibliografia

[Baldan et al.,] Baldan, P., Cocco, N., De Nes, F., Llabres, M., Marin, A., and Simeoni, M. Mpath2pn: Translating metabolic pathways into petri nets. In *In Proc. of Int. Work. on Biological Processes and Petri Nets*, Newcastle upon Tyne, UK.

[Baldan et al., 2010] Baldan, P., Cocco, N., Marin, A., and Simeoni, M. (2010). Petri nets for modelling metabolic pathways: a survey. 9(4):955–989.

[Chen and Hofestadt, 2004] Chen, M. and Hofestadt, R. (2004). Web-based information retrieval system for the prediction of metabolic pathways. *IEEE Trans Nanobioscience*, 3(3):192–9.

[Clemente et al., 2005] Clemente, J., Satou, K., and Valiente, G. (2005). Reconstruction of phylogenetic relationships from metabolic pathways based on the enzyme hierarchy and the gene ontology. *Genome Inform*, 16(2):45–55.

[Deville et al., 2003] Deville, Y., Gilbert, D., Van Helden, J., and Wodak, S. (2003). An overview of data models for the analysis of biochemical pathways. *Briefings in Bioinformatics*, 4:246–259.

[Dingle et al., 2009] Dingle, N. J., Knottenbelt, W. J., and Suto, T. (2009). Pipe2: a tool for the performance evaluation of generalised stochastic petri nets. *SIGMETRICS Perform. Eval. Rev.*, 36:34–39.

[Feng and Doolittle, 1987] Feng, D. F. and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution*, 25(4):351–360.

[Ferhat and Kahveci, 2010] Ferhat, A. and Kahveci, T. (2010). Submap: Aligning metabolic pathways with subnetwork mappings. In Berger, B., editor, *Research in Computational Molecular Biology, 14th Annual International Conference, RECOMB 2010, Lisbon, Portugal, April 25-28, 2010.*

Proceedings, volume 6044 of *Lecture Notes in Computer Science*, pages 15–30. Springer.

[Forst et al., 2006] Forst, C., Flamm, C., Hofacker, I., and Stadler, P. (2006). Algebraic comparison of metabolic networks, phylogenetic inference, and metabolic innovation. *BMC Bioinformatics*, 7(1):67.

[Hardy and Robillard, 2004] Hardy, S. and Robillard, P. N. (2004). Pn: Modeling and simulation of molecular biology systems using petri nets: modeling goals of various approaches. *J Bioinform Comput Biol*, 2004:2–4.

[Hattori et al.,] Hattori, M., Okuno, Y., Goto, S., and Kanehisa, M. Heuristics for chemical compound matching.

[Hattori et al., 2003] Hattori, M., Okuno, Y., Goto, S., and Kanehisa, M. (2003). Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *Journal of the American Chemical Society*, 125(39):11853–11865.

[Heiner and Koch, 2004] Heiner, M. and Koch, I. (2004). Petri net based model validation in systems biology. In *In 25th International Conference on Application and Theory of Petri Nets*, pages 216–237. Springer.

[Heiner et al., 2001] Heiner, M., Koch, I., and Voss, K. (2001). Analysis and simulation of steady states in metabolic pathways with petri nets.

[Heymans and Singh, 2002] Heymans, M. and Singh, A. K. (2002). Deriving phylogenetic trees from the similarity analysis of metabolic pathways.

[Hong et al., 2004] Hong, S. H., Kim, T. Y., and Lee, S. Y. (2004). Phylogenetic analysis based on genome-scale metabolic pathway reaction content. *Appl Microbiol Biotechnol*, 65(2):203–10.

[KEGG, 2012] KEGG, K. U. B. C. (2012). Kegg pathway database. `http://www.genome.jp/kegg/pathway.html`.

[Li et al., 2008] Li, Y., de Ridder, D., de Groot, M., and Reinders, M. (2008). Metabolic pathway alignment between species using a comprehensive and flexible similarity measure. *BMC Systems Biology*, 2(1):111.

[Liao et al., 2002] Liao, L., Kim, S., and Tomb, J.-f. (2002). Genome comparisons based on profiles of metabolic pathways. In *Proceedings of the 6th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES '02*, pages 469–476.

[Lo et al., 2004] Lo, E., Yamada, T., Tanaka, M., Hattori, M., Goto, S., Chang, C., and Kanehisa, M. (2004). A method for customized cross-species metabolic pathway comparison. *In Proc. of Genome Informatics*.

[Needleman and Wunsch, 1970] Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48:443–453.

[Peterson, 1981] Peterson, J. (1981). *Petri Net Theory and the Modelling of Systems*. Prentice-Hall.

[Pinter et al., 2005] Pinter, R. Y., Rokhlenko, O., Yeger-Lotem, E., and Ziv-Ukelson, M. (2005). Alignment of metabolic pathways. *Bioinformatics*, 21:3401–3408.

[Plummer, 2009] Plummer, M. D. (2009.). *Matching theory /*. AMS Chelsea Pub.„ Providence, R.I. :. Originally published: Amsterdam ; New York : North-Holland, 1986.

[Reddy et al., 1993] Reddy, V. N., Mavrovouniotis, M. L., and Liebman, M. N. (1993). Petri net representations in metabolic pathways. In *Proceedings of the 1st International Conference on Intelligent Systems for Molecular Biology*, pages 328–336. AAAI Press.

[SBML, 2012] SBML (2012). Systems biology markup language. `http://sbml.org`.

[Shaw et al., 2004] Shaw, O., Koelmans, A., Steggles, J., and Wipat, A. (2004). Applying petri nets to systems biology using XML technologies. Technical report, University of Newcastle upon Tyne.

[Smith and Waterman, 1981] Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195 – 197.

[Tohsato, 2007] Tohsato, Y. (2007). A method for species comparison of metabolic networks using reaction profile. *Information and Media Technology*, page 109–114.

[Tohsato et al., 2000] Tohsato, Y., Matsuda, H., and Hashimoto, A. (2000). A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB '00*, pages 376–383.

[Wernicke and Rasche, 2007] Wernicke, S. and Rasche, F. (2007). Simple and fast alignment of metabolic pathways by exploiting local diversity. *Bioinformatics*, 23(15):1978–1985.

[Zevedei-Oancea and Schuster, 2003] Zevedei-Oancea, I. and Schuster, S. (2003). Topological analysis of metabolic networks based on petri net theory. *In Silico Biology*, 3:29.