Αλέξανδρος Ξιάρχος
st1059619@ceid.upatras.gr
1059619

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ · ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΕΙΣΑΓΩΓΗ ΣΤΗ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ: XML ΚΑΙ ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ

ΔΕΥΤΕΡΟ ΣΥΝΟΛΟ ΑΣΚΗΣΕΩΝ · 2023 – 2024

ΠΕΡΙΕΧΟΜΕΝΑ

1 ΕΙΣΑΓΩΓΗ

Η βιοπληροφορική έχει αναδειχθεί σαν ένα κομβικό επιστημονικό πεδίο ανάμεσα στη βιολογία και την επιστήμη των υπολογιστών. Χρησιμοποιεί υπολογιστικά εργαλεία για τη μελέτη και την κατανόηση βιολογικών δεδομένων όπως το DNA και οι πρωτεϊνες, μια διαδικασία που η ραγδαία πρόοδος των επιστημών κατέστησε απαραίτητη. Η βιοπληροφορική πλέον έχει εξελιχθεί σε ένα αναγκαίο εργαλείο για ερευνητικούς σκοπούς, στην ανακάλυψη νέων φαρμάκων, στην εξατομικευμένη και προληπτική ιατρική, στη γονιδιακή θεραπεία, στη βελτιώση της καλλιέργειας κ.α. Η χρήση της βοηθάει στην επέκταση της γνώσης πολύ πιο αποτελεσματικά και με μεγαλύτερη ακρίβεια.

Ο συγκεκριμένος τομέας ήρθε στο προσκήνιο με την ανακάλυψη του ανθρώπινου γονιδιώματος, κάτι που οι παραδοσιακές μέθοδοι ανάλυσης δεδομένων ήταν ανεπαρκείς για να χειριστούν τον τεράστιο όγκο των πληροφοριών που παραγόνταν.

Μια μεγάλη πρόκληση στη βιοπληροφορική είναι η ανάγκη για τυποποίηση των τρόπων αναπαράστασης και ανταλλαγής πολύπλοκων βιολογικών δεδομένων. Εδώ είναι που η XML (eXtensive Markup Language) μπαίνει στο προσκήνιο. Πρόκειται για μια γλώσσα σήμανσης αρκετά ευέλικτη και ισχυρή για την αναπαράσταση ιεραρχικών σχέσεων (hierarchical relationships), κάτι αρκετά κοινότυπο στη μελέτη βιολογικών δεδομένων.

Η σύνδεση μεταξύ XML και βιοπληροφορικής είναι τόσο διαδεδομένη που έχει οδηγήσει στην ανάπτυξη διάφορων ευρέως χρησιμοποιούμενων προτύπων στο τομέα που θα αναλυθούν στη συνέχεια. Καθώς ο τομέας της βιοπληροφορικής συνεχίζεται να εξελίσσεται, ο ρόλος των γλωσσών σήμανσης όπως η XML και των προτύπων που αυτή δημιουργεί γίνεται ολοένα και πιο σημαντικός για την ανταλλαγή και ενοποίηση δεδομένων και την παραγωγικότητα της επιστημονικής έρευνας.

2 ΑΝΑΣΚΟΠΗΣΗ ΤΩΝ ΑΡΘΡΩΝ

2.1 ΠΡΟΣΕΓΓΙΣΕΙΣ ΜΕ ΒΑΣΗ ΤΗΝ XML

Το άρθρο "*XML-based approaches for the integration of heterogeneous bio-molecular data*" των Mesiti, Jimenez-Ruiz κ.α. πραγματεύεται κάποιες προσεγγίσεις για την αναπαράσταση, ενσωμάτωση και διαχείριση βιολογικών δεδομένων με τη χρήση της XML γλώσσας και άλλων παρεμφερών υλοποιήσεων. Επιπλέον, παρουσιάζεται μια νέα προσέγγιση για τη διαχείριση ετερογενών βιολογικών δεδομένων μέσω της XML. [3]

2.1.1 Χρησιμότητα XML

Η XML έχει αναδειχθεί ως την πιο αποτελεσματική πρόταση για την αναπαράσταση δομημένων πληροφοριών, μιας και επιτρέπει την εύκολη επέκταση και τροποποίηση, κάτι βολικό μιας και καθημερινά δημιουργούνται νέα βιολογικά δεδομένα. Υποστηρίζεται από γλώσσες ερωτημάτων (query languages) όπως η XPath και XQuery, δίνοντας τη δυνατότητα για άμεση εξόρυξη των πληροφοριών.

Χρησιμοποιεί μια ιεραρχική δόμηση της πληροφορίας με στοιχεία (XML Elements), χαρακτηριστικά (XML attributes) και κείμενο (XML text content). Κάθε στοιχείο μπορεί να αναπαριστά κάποια συγκεκριμένη βιολογική οντότητα (πχ DNA, RNA, πρωτεΐνη) και μπορεί να περιλαμβάνει εμφωλευμένα στοιχεία για συσχετιζόμενα χαρακτηριστικα. Αυτή η ιεραρχική δόμηση αναπαριστά με σαφήνεια πολύπλοκες βιολογικές σχέσεις.

2.1.2 Βιολογικοί τύποι δεδομένων

Το άρθρο παρουσιάζει κάποιους τύπους βιολογικών δεδομένων. Για παράδειγμα:

- **Δεδομένα πρωτοταγών πρωτεϊνών**: περιλαμβάνουν δεδομένα νουκλεοτιδικών αλληλουχιών,

φιλοξενούνται σε βάσεις δεδομένων όπως GenBank και EMBL.

- **Δεδομένα πρωτεϊνών**: βάσεις δεδομένων όπως SWISSPROT και TREMBL περιέχουν πληροφορίες για πρωτεϊνικές αλληλουχίες, αναπαρίστανται εύκολα σε XML.
- **Motif (μοτίβα) και πρωτεϊνικές περιοχές**: προσδιορίζονται μέσω μεθόδων αναγνώρισης προτύπων που εφαρμόζονται σε δεδομένα πρωτοταγών πρωτεϊνών, αναπαρίστανται με μια περιγραφή του μοτίβου, βιβλιογραφικές πληροφορίες κ.α.

2.1.3 Αναπαράσταση βιολογικών δεδομένων με τη χρήση XML

Έχουν χρησιμοποιηθεί αρκετές γλώσσες βασισμένες στην XML, ειδικά για την αναπαράσταση διαφορετικών τύπων βιολογικών δεδομένων. Για παράδειγμα:

2.1.3.1 Bioinformatic Sequence Markup Language (BSML)

Γλώσσα σχεδιασμένη για να περιγράφει αλληλουχίες όπως DNA, RNA και πρωτεϊνικές αλληλουχίες. Ένα BSML αρχείο περιλαμβάνει πληροφορίες για το πώς τα γονιδιώματα κωδικοποιούνται, ανακτώνονται και εμφανίζονται.

2.1.3.2 Protein Markup Language - ProXML

Χρησιμοποιείται για την αναπαράσταση πρωτεϊνικών αλληλουχιών. Περιλαμβάνει ένα identity section που περιέχει την περιγραφή των πρωτεϊνών και ένα data section που περιέχει ιδιότητες από αυτές τις πρωτεΐνες.

2.1.3.3 ΠΡΟΕΚΤΑΣΗ: RNA Markup Language - RNAML

Γλώσσα που σχεδιάστηκε με σκοπό να διευκολύνει την ανταλλαγή RNA πληροφοριών μεταξύ διαφορετικών λογισμικών βιοπληροφορικής. [4] Μέχρι πρότινος, κάθε εργαστήριο ανέπτυσσε το δικό του λογισμικό με τους δικούς του τύπους αρχείων για την ανάγνωση και την εγγραφή της βιοπληροφορίας. Επομένως, κατέστη αναγκαία η δημιουργία μιας τυποποιημένης σύνταξης της RNA πληροφορίας, η οποία έχει σκοπό να αυξήσει την αποτελεσματικότητα στην κοινότητα των βιολόγων.

Οι προηγούμενες προσπάθειες για τυποποίηση της βιολογικής πληροφορίας περιλαμβάνουν την BIOpolymer Markup Language (BIOML) γλώσσα σήμανσης, η οποία αναπτύχθηκε το 1999 από την ProteoMaterics. Περιλαμβάνει ένα διευρυμένο framework για τον καθορισμό μοριακών οντοτήτων, και ενώ περιλάμβανε κάποιες πληροφορίες για το RNA, εστιάζονταν περισσότερο στη γονιδιακή του πλευρά (θέσεις έναρξης και παύσης της μεταγραφής και γενετικές τροποποιήσεις). Δεν κάλυπτε επαρκώς πληροφορίες για τη δομή του RNA, κάτι ερευνητικά κρίσιμο. Επομένως, δεν κάλυπταν τις απαιτήσεις που έθετε η επιστημονική κοινότητα που μελετούσε το RNA, οδηγώντας στην ανάπτυξη της RNAML.

Η RNAML βασίζεται πάνω στο XML. Υπάρχει η δυνατότητα δημιουργίας ενός Document Type Definition (DTD) το οποίο καθορίζει τη δομή του εγγράφου, τα ονόματα και τον τύπο των στοιχείων και τη ιεραρχική δομή τους, κάτι που διασφαλίζει τη συνέπεια και τη συμμόρφωση στο πώς αναπαρίσταται το RNA. Επίσης, αναπαριστά πληροφορία για την αλληλεπίδραση πολλαπλών μορίων RNA, την απόσταση τους, τη σύζευξη των βάσεών τους, και οποιαδήποτε άλλη σχέση έχουν μεταξύ τους. Τέλος, πέρα από δυνατότητες για σχολιασμό και documentation σε κάθε στοιχείο, είναι δυνατή η ομαδοποίηση των εμφανίσεων του ίδιου λειτουργικού RNA σε διαφορετικούς οργανισμούς, επιτρέποντας την αναπαράσταση ευθυγραμμίσεων και κοινών δομικών συστατικών.

2.1.3.4 ΠΡΟΕΚΤΑΣΗ: System Biology Markup Language (SBML)

Δημιουργήθηκε στα πλαίσια του ERATO Kitano Systems Biology Project για να διευκολύνει την ανταλλαγή μοντέλων μεταξύ διαφορετικών εργαλείων προσομοίωσης και ανάλυσης. [1]

Ένα SBML μοντέλο περιλαμβάνεται από το Διαμέρισμα (Compartment), έναν καθορισμένο χώρο όπου συμβαίνουν οι αντιδράσεις όπως ένα κύτταρο ή ένα οργανίδιο, ένα Είδος (Species), οι χημικές οντότητες που συμμετέχουν στις αντιδράσεις όπως τα ιόντα ή τα μόρια, η Αντίδραση (Reaction), η διαδικασία σχηματισμού μεταξύ των ειδών, η Παράμετρος (Parameter), όπου αναπαριστά ποσότητες με συμβολικά ονόματα τοπικά η καθολικά, Ορισμοί Μονάδων (Unit Definitions), για τον προσδιορισμό των μονάδων που χρησιμοποιούνται στο μοντέλο, και τέλος οι Κανόνες (Rules), μαθηματικές εκφράσεις που ορίζουν τις τιμές των παραμέτρων ή θέτουν περιορισμούς στο μοντέλο.

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <sbml xmlns="http://www.sbml.org/sbml/level1"
3        level="1" version="2">
4    <model name="gene_network_model">
5      <listOfUnitDefinitions>
6          ...
7      </listOfUnitDefinitions>
8      <listOfCompartments>
9          ...
10     </listOfCompartments>
11     <listOfSpecies>
12         ...
13     </listOfSpecies>
14     <listOfParameters>
15         ...
16     </listOfParameters>
17     <listOfRules>
18         ...
19     </listOfRules>
20     <listOfReactions>
21         ...
22     </listOfReactions>
23   </model>
24 </sbml>
```

Σχήμα 2.1: Σκελετός από τον ορισμό ενός μοντέλου [1]

2.1.3.5 ΠΡΟΕΚΤΑΣΗ: Cell Markup Language (CellML)

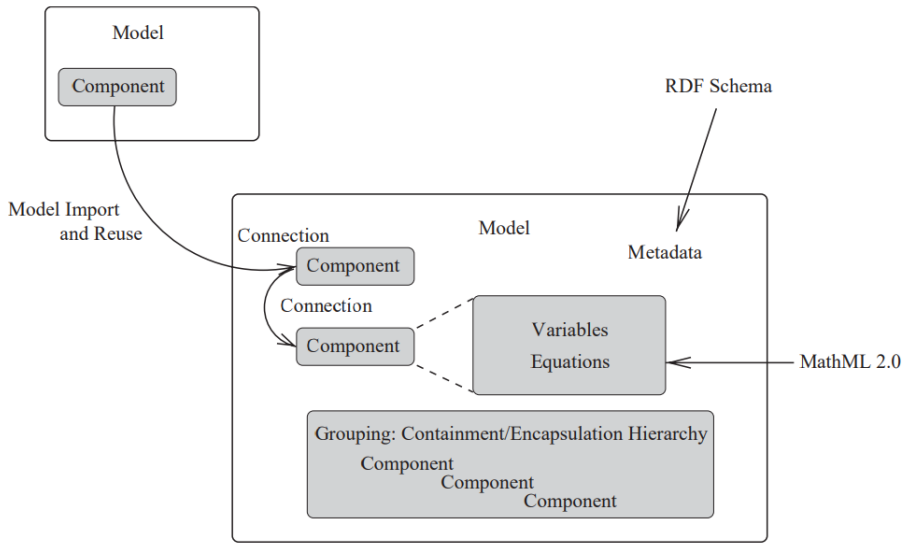
Το CellML προσφέρει μια σαφή μέθοδο ορισμού μοντέλων κυτταρικής λειτουργίας, σε ένα πιο γενικό πλαίσιο σε σχέση με τα προηγούμενα. [2] Το βάθος στο οποίο μπορεί το CellML να αναπαραστήσει τις έννοιες επικαλύπτει γλώσσες όπως η SBML, με τη διαφορά ότι η SBML βασίζεται περισσότερο στην περιγραφή βιοχημικών αντιδράσεων, χάνοντας πληροφορία για τη δομή των μοντέλων.

Η CellML από την αρχή σχεδιάστηκε για να υποστηρίξει μοντέλα μεγάλης κλίμακας, επιτρέπεται (λόγω της XML βάσης της) η ανεξάρτητη κατασκευή μοντέλων και τμημάτων και η ενσωμάτωσή τους σε ένα μεγαλύτερο μοντέλο, και παρέχει τρόπους για την απόκρυψη low-level πληροφοριών ώστε να μη συγχέονται με το υψηλότερο επίπεδο αναπαράστασης του μοντέλου.

2.1.4 Διαχείριση ετερογενών βιολογικών δεδομένων

Οι βιολόγοι συνήθως χρησιμοποιούν διαφορετικές βάσεις δεδομένων, η κάθε μία με το δικό της σχεδιασμό της πληροφορίας, που καθιστά χρονοβόρα τη ανάκτηση πληροφορίας. Επομένως, είναι αυξημένη η ανάγκη για πρόσβαση σε μια ομογενοποιημένη βάση δεδομένων, κάτι που δεν είναι πάντα εύκολο να επιτευχθεί λόγω της ετερογένειας της πληροφορίας.

Λύση σε αυτό είναι η γλώσσα XML, που προφέρει έναν τρόπο για τη συντακτική ενσωμάτωση των δεδομένων, αν και στερείται των μεθόδων με τους οποίους μπορεί να επιτευχθεί αυτή η ενσωμάτωση. Τέτοιες



Σχήμα 2.2: Διάγραμμα με το σκελετό ενός CellML μοντέλου [2]

μέθοδοι ονομάζονται αρχιτεκτονικές ενσωμάτωσης (integration architectures) και χωρίζονται στις Data warehouse, Mediator-based, Service oriented και Peer-based αρχιτεκτονικές.

Το δεύτερο μέρος του άρθρου αναλύει τη χρήση αυτών των αρχιτεκτονικών σε συνδυασμό με την XML για την ενσωμάτωση των δεδομένων.

Βιβλιογραφία

[1] M. Hucka κ.ά. “The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models”. Στο: *Bioinformatics* 19.4 (Μαρ. 2003), σσ. 524–531. DOI: 10.1093/bioinformatics/btg015.

[2] Catherine M. Lloyd, Matt D.B. Halstead και Poul F. Nielsen. “CellML: Its future, present and past”. Στο: *Progress in Biophysics and Molecular Biology* 85.2–3 (Ιούν. 2004), σσ. 433–450. DOI: 10.1016/j.pbiomolbio.2004.01.004.

[3] Marco Mesiti κ.ά. “XML-based approaches for the integration of heterogeneous bio-molecular data”. Στο: *BMC Bioinformatics* 10.S12 (Οκτ. 2009). DOI: 10.1186/1471-2105-10-s12-s7.

[4] ALLISON WAUGH κ.ά. “RNAML: A standard syntax for exchanging RNA information”. Στο: *RNA* 8.6 (Ιούν. 2002), σσ. 707–717. DOI: 10.1017/s1355838202028017.