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. [5]

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. [6] 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. [2]

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 [2]

2.1.3.5 ΠΡΟΕΚΤΑΣΗ: Cell Markup Language (CellML)

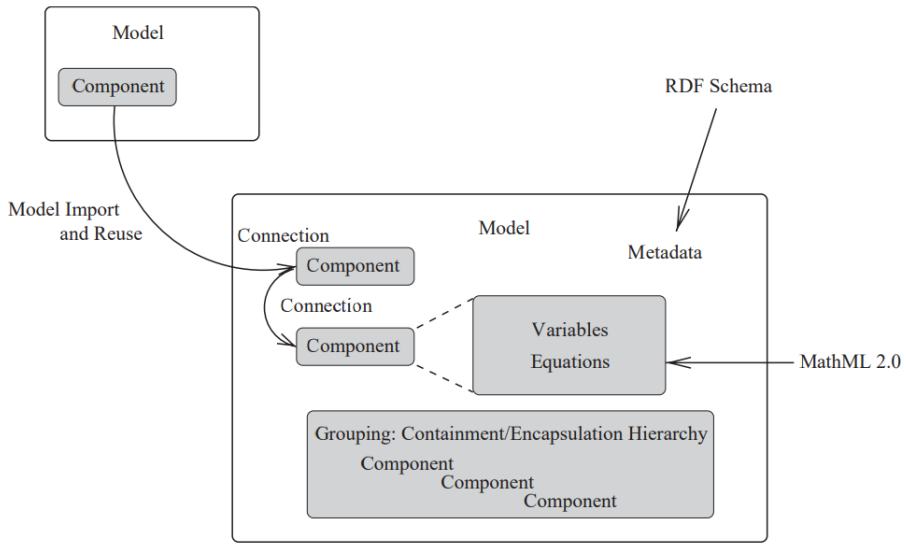
Το CellML προσφέρει μια σαφή μέθοδο ορισμού μοντέλων κυτταρικής λειτουργίας, σε ένα πιο γενικό πλαίσιο σε σχέση με τα προηγούμενα. [4] Το βάθος στο οποίο μπορεί το CellML να αναπαραστήσει τις έννοιες επικαλύπτει γλώσσες όπως η SBML, με τη διαφορά ότι η SBML βασίζεται περισσότερο στην περιγραφή βιοχημικών 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 την 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 [4]

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.

2.1.4.1 ΠΡΟΕΚΤΑΣΗ: Data warehouse συστήματα

Η Data warehouse αρχιτεκτονική ενσωματώνει δεδομένα από διαφορετικές βάσεις δεδομένων σε μια, καταφέρνοντας μια υψηλότερου βαθμού ομογενοποίηση και χωρίς να χρειάζονται συχνές ανανεώσεις. Παραδείγματα τέτοιων συστημάτων είναι τα εξής:

DWARF

Πρόκειται για ένα data warehouse σύστημα που σχεδιάστηκε για την ανάλυση μεγάλων πρωτεϊνικών οικογενειών. Ενσωματώνει δεδομένα σχετικά με την αλληλουχία, δομή και τον χαρακτηρισμό πρωτεϊνών, συνδυάζοντας δεδομένα από διαφορετικές δημόσιες βάσεις δεδομένων όπως GenBank, ExPDB, κα.

Το σχεσιακό του μοντέλο δεδομένων αναπτύχθηκε στο Firebird, ένα ανοιχτού κώδικα σύστημα διαχείρισης σχεσιακών βάσεων SQL, και είναι οργανωμένο σε τρία μεγάλα τμήματα που αντιπροσωπεύουν διαφορετικές οντότητες: την πρωτεϊνη (περιγράφει τη βιοχημική λειτουργία, τον οργανισμό προέλευσης και την ταξινόμηση των πρωτεϊνών), την αλληλουχία των πρωτεϊνών (σχολιασμός συγκεκριμένων θέσεων, λεπτομέρειες για μεταλλάξεις) και δομή πρωτεΐνης (δεδομένα που σχετίζονται με τις δευτερογενείς και τριτογενείς δομές της πρωτεΐνης). [1]

BioWarehouse

Πρόκειται για ένα toolkit ανοιχτού κώδικα που έχει σχεδιαστεί για τη διευκόλυνση της διασύνδεσης διαφορετικών βάσεων δεδομένων βιοπληροφορικής. Χρησιμοποιεί ως relational database managers τη MySQL και την Oracle, και επιτρέπει την ομαλή σύνδεση διαφορετικών βάσεων δεδομένων επιτρέποντας αποτελεσματικά queries και την εξόρυξη δεδομένων. [3]

Περιλαμβάνονται εργαλεία σε C και σε Java που κάνουν συντακτική ανάλυση (parsing) και κανονικοποιούν τα δεδομένα για να μειώσουν την ετερογένεια, ενώ έχει σχεδιαστεί ώστε να επιτρέπεται η κλιμάκωση για πολλά terabytes δεδομένων Όλα αυτά κάνουν το BioWarehouse ένα χρήσιμο εργαλείο με

πολλές πρακτικές εφαρμογές, όπως για παράδειγμα για τον προσδιορισμό κενών σε αλληλουχίες.

Atlas

Data warehouse σύστημα που αποθηκεύει και ενοποιεί διαφορετικούς τύπους βιολογικών δεδομένων. Βασίζεται στη χρήση της SQL γλώσσας που εφαρμόζεται σε API κλήσεις τριών γλωσσών (C++, Java, Perl), οι οποίες διαβάζουν πληροφορίες από άλλες βάσεις δεδομένων (GenBank, RefSeq, UniProt κα) στη βάση δεδομένων του Atlas. Επίσης, περιλαμβάνει κάποια εργαλεία που χρησιμοποιούνται για την εξόρυξη δεδομένων. [**Atlas**]

Biozone

Ενοποιημένη πηγή για DNA αλληλουχίες, πρωτεΐνες κα, που ενσωματώνει μοντέλα γράφων και ιεραρχικές κλάσεις για την αναπαράσταση και την κατηγοριοποίηση βιολογικών οντοτήτων.

cPath

Λογισμικό βάσης δεδομένων ανοιχτού κώδικα για τη συλλογή, αποθήκευση και αναζήτηση δεδομένων βιολογικών μονοπατιών. Τα δεδομένα μπορούν και προβάλλονται σε browser ή να εξαχθούν μέσω API που βασίζεται σε XML, κάτι που επιτρέπει τη χρήση του σε third-party εφαρμογές φτιαγμένες για την οπτικοποίηση και ανάλυση μονοπατιών.

Aspect	DWARF	BioWareh.	Atlas	Biozone	CPath
BioData	Sequences	All Types	Genes	All Types	AllTypes
Instantiation			Materialized		
Integration			Common Storage/Access		
Global View		LAV		GAV (I)	LAV
Global Model		Relational		Graph	RDF/OWL
Query Model		SQL		SQL/AdHoc	SPARQL
Semantics	-	Thesaurus	-	-	Ontologies
Scalability	Low	Medium	Medium	Medium	Medium

Σχήμα 2.3: Σύγκριση των Data Warehouse συστημάτων [5]

2.1.4.2 ΠΡΟΕΚΤΑΣΗ: Mediator-based συστήματα

Σε αυτή την αρχιτεκτονική, οι διάφορες βάσεις δεδομένων διατηρούν την αυτονομία τους και τα mediator-based συστήματα δρουν ως μεσάζοντες. Ο στόχος είναι η δημιουργία μιας ενοποιημένης προβολής των δεδομένων (global view) χωρίς να είναι απαραίτητη η φυσική μεταφορά των δεδομένων σε μια βάση. Κάθε ξεχωριστή βάση απαιτεί τον ορισμό ενός wrapper, ο οποίος θα μετατρέψει τη μορφή των δεδομένων τους (από/σε XML για παράδειγμα).

Τα κύρια πλεονεκτήματα της συγκεκριμένης αρχιτεκτονικής είναι ότι τα δεδομένα είναι πάντα ανανεωμένα (up-to-date), δεν υπάρχουν διπλότυπα και είναι ευκολότερη η ενσωμάτωση νέων πηγών δεδομένων. Το μεγάλο μειονέκτημα προφανώς είναι ο χειροκίνητος καθορισμός του wrapper που απαιτείται για την ενσωμάτωση των δεδομένων, αν και έχουν προταθεί κάποιες τεχνικές αυτοματοποίησης. Παραδείγματα τέτοιων συστημάτων είναι τα εξής:

Ontofusion

Σύστημα οντολογίας που βασίζεται σε δύο διεργασίες: χαρτογράφηση (mapping) και ενοποίηση (unification).

Η χαρτογράφηση είναι μια ημιαυτόματη διαδικασία που χρησιμοποιεί οντολογίες (virtual schemas) για τη σύνδεση των εξωτερικών βάσεων δεδομένων. Χρησιμοποιούνται τρεις μέθοδοι για τη δημιουργία των

οντολογιών: η top-down (χρησιμοποιώντας μια υπάρχουσα οντολογία όπως UML), η bottom-up (χτίζοντας μια νέα οντολογία) και ο συνδυασμός τους.

Οι οντολογίες αυτές συνενώνονται σε ένα ξεχωριστό "global schema" όπου πλέον είναι ομογενοποιημένες. [**Ontofusion**]

TAMBIS

Το TAMBIS (Transparent Access to Multiple Bioinformatics Information Sources) χρησιμοποιεί την Tambis Οντολογία (TAO), ως ένα κοινό framework για την ενσωμάτωση διαφορετικών βάσεων δεδομένων. Υπάρχουν δύο εκδοχές του, μια unlinked εφαρμογή που επιτρέπει στον χρήστη να πλοηγηθεί σε ένα μοντέλο με 1800 έννοιες της βιολογίας και ένα μοντέλο που είναι συνδεδεμένο με εξωτερικές βάσεις δεδομένων. [**TAMBIS**]

Aspect	Ontofusion	TAMBIS	Biomed.	WS	P2P
BioData	Genes	All types	Genes	All Types	All Types
Instantiation			Virtual		
Integration			Common Access		
Global View	GAV (S/I)	GAV (S)	LAV	LAV	N.A.
Global Model	RDF/OWL		XML	RDF/OWL	XML
Query Model	Boolean	CPL	XQuery	SPARQL	XQuery
Semantics	Ontologies		-	-	-
Scalability	Medium	Low	Medium	High	High

Σχήμα 2.4: Σύγκριση των Mediator-based συστημάτων [5]

2.1.4.3 Service-oriented συστήματα

Η Service-oriented αρχιτεκτονική προσφέρει μια τυποποιημένη μέθοδο για την ενσωμάτωση και των δεδομένων και του λογισμικού, θεωρώντας τα ως υπηρεσίες. Έτσι οι εφαρμογές θα τις συνδυάσουν για να υλοποιήσουν τις προβλεπόμενες εργασίες τους.

2.1.4.4 Peer-based συστήματα

Προσφέρουν μια αποκεντρωμένη προσέγγιση για την ενσωμάτωση δεδομένων μεταξύ διαφορετικών πηγών δεδομένων σε ένα δίκτυο. Η αποκέντρωση προσφέρει μεγαλύτερη ευελιξία και επεκτασιμότητα, ενώ δεν είναι απαραίτητη η δημιουργία μιας κεντρικής οντολογίας στην οποία πρέπει να μετατραπούν όλα τα δεδομένα. Από την άλλη, αυτά τα συστήματα δεν είναι τόσο αποδοτικά, καθώς η πολυπλοκότητα των δεδομένων είναι αυξημένη.

2.1.5 Παράμετροι για την ενοποίηση των δεδομένων

Κάποιες από τις παραμέτρους που επηρεάζουν την αρχιτεκτονική που θα χρησιμοποιήσουμε για την ενοποίηση των δεδομένων είναι **οι τύποι των δεδομένων** (όλα βασίζονται στο XML, αλλά διαφορετικά συστήματα ενοποίησης εστιάζουν σε διαφορετικούς τύπους δεδομένων όπως αλληλουχίες, γονιδιακές εκφράσεις κτλ), το **global model** (η μορφή της αναπαράστασης: relation-based (SQL), tree-based (XML), graph-based (RDF)), το **query model** (οι γλώσσες που χρησιμοποιούνται για την πρόσβαση στα δεδομένα όπως SQL, XQuery κτλ), η **επεκτασιμότητα** κ.α.

2.1.6 Εξειδικευμένα θέματα στην ενοποίηση XML δεδομένων

Το άρθρο κάνει αναφορά σε κάποια επιπλέον ζητήματα που αφορούν την ενοποίηση των δεδομένων που βασίζονται στο XML.

Τίθονται θέματα που αφορούν την ασφάλεια των δεδομένων και την ιδιωτικότητα, την εξέλιξη των

δεδομένων λόγω της δυναμικής φύσης τους, την αποτελεσματικότητα των ερωτήσεων (queries) που θέτουμε όπως επίσης και την έλλειψη -για την ώρα- μιας τυποποιημένης αρχιτεκτονικής που να εφαρμόζεται καθολικά.

Σε κάθε περίπτωση, το XML έχει ξεκάθαρα επιτύχει ως τη συντακτική κόλλα που συνδέει διάφορες πηγές με βιολογικά δεδομένα. Το αρνητικό είναι πως έχει δημιουργήσει μια μεγάλη ποικιλία διαφορετικών μορφών δεδομένων, κάτι που καθιστά δύσκολη την αποτελεσματική ενσωμάτωσή τους.

2.2 EDAM

Το άρθρο "*EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics and formats*" των Ison, Kalas, Jonassen κ.α. πραγματεύεται την οντολογία EDAM, μια οντολογία που έχει σχεδιαστεί για την κατηγοροποίηση πράξεων και τύπων δεδομένων στην βιοπληροφορική. [**EDAMpaper**] Η EDAM (EMBRACE Data and Models) είναι μια οντολογία από έννοιες που είναι διαδεδομένες στην ανάλυση βιοεπιστημονικών δεδομένων, περιλαμβάνει πράξεις (operations), τύπους δεδομένων (data types), data identifiers και data formats, όπως επίσης και συνώνυμα, συναφείς όρους, ορισμούς και άλλες πληροφορίες, όλα οργανωμένα με ιεραρχικό τρόπο. [**EDAMsite**]

2.2.1 ΠΡΟΕΚΤΑΣΗ: Τι είναι η οντολογία;

Γενικά οντολογία είναι μια επίσημη αναπαράσταση γνώσης σε ένα συγκεκριμένο πεδίο. Ορίζει ένα σύνολο εννοιών και κανόνων όπως επίσης και τις μεταξύ τους σχέσεις σε ένα δομημένο format που μπορεί να διαβαστεί εύκολα και από της μηχανές. Υπάρχει ιεράρχηση στις έννοιες μέσω της δημιουργίας κλάσεων, ιδιότητες (attributes) από αυτές τις έννοιες, σχέσεις (relationships) μεταξύ τους όπως επίσης και παραδείγματα (instances) από τις έννοιες. Μια οντολογία αναπαρίσταται σε γλώσσες σήμανσης όπως η OWL (Web Ontology Language) ή RDF (Resource Description Framework XML), γλώσσες που έχουν βασιστεί πάνω στην XML. [**OntologyWiki**] [**BioOntologies**]

2.2.2 Πώς δημιουργήθηκε η ανάγκη για να δημιουργηθεί η οντολογία EDAM;

Σε μια εποχή ταχείας ανάπτυξης της βιολογίας και της πληροφορικής με αποτέλεσμα την αύξηση των πληροφοριακών αναγκών, μέχρι πρότινος δεν υπήρχε μια ολοκληρωμένη οντολογία με σωστές ταξινομημένες πληροφορίες κατά ένα τρόπο που να υποστηρίζονται πράξεις μεγάλης κλίμακας. Τα λεξικά και οι οντολογίες που είχαν δημιουργηθεί ως τότε δεν κάλυπταν πλήρως τις ανάγκες που απαιτούσε η επιστήμη της βιοπληροφορικής.

2.2.3 Τι προσφέρει η οντολογία EDAM;

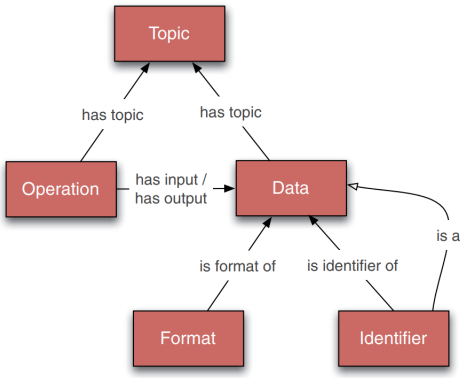
Η οντολογία EDAM σχεδιάστηκε με γνώμονα να είναι σχετική με τις προβλεπόμενες εφαρμογές της στη βιοπληροφορική. Για να επιτευχθεί αυτό, μελετήθηκαν οι υπάρχουσες οντολογίες και εργαλεία (myGrid ontology (2007), εργαλεία από την EMBRACE, BioMoby Object Ontology (2001)). Ακολουθούνται αρχές όπως η λογική συνέπεια (να μην υπάρχουν αντιφάσεις στην οντολογία), σαφές σημασιολογικό πεδίο (είναι καθορισμένα τα όρια του τι καλύπτει) όπως επίσης και η δυνατότητα για μελλοντικές προεκτάσεις.

2.2.4 Υλοποίηση EDAM

Η οντολογία χρησιμοποιεί URIs (Uniform Resource Identifiers) για τη μοναδική αναπαράσταση των εννοιών της, με τη μορφή: `http://e-damontology.org/<subontology>_<localId>`.

Υπάρχουν τέσσερις διαφορετικές υποοντολογίες: **Operation** (συνάρτηση με εισόδους και εξόδους, πχ πρόβλεψη δομής RNA), **Data** (πληροφορία, πχ αλληλουχίες) και τη δική του υπακολουθία **Identifier**, **Topic** (περιγράφει κατηγορίες, πχ «Ανάλυση αλληλουχιών») και **Format** (μορφές δεδομένων, πχ FASTQ). Έτσι, η

κλάση «Sequence record» για παράδειγμα αναπαρίσταται ως `http://edamontology.org/data_0849`. Οι σχεσιακές σχέσεις και οι υπόλοιπες ιδιότητες ορίζονται με τη μορφή: `http://edamontology.org/<id>`.



Σχήμα 2.5: Σχέσεις μεταξύ των υποοντολογιών [**EDAMpaper**]

2.3 METATRON

2.4 ΤΕΛΕΥΤΑΙΟ PAPER

Βιβλιογραφία

[1] Markus Fischer κ.ά. “Dwarf – a data warehouse system for analyzing protein families”. Στο: *BMC Bioinformatics* 7.1 (Νοέ. 2006). DOI: 10.1186/1471-2105-7-495.

[2] 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.

[3] Thomas J Lee κ.ά. “BioWarehouse: A bioinformatics database warehouse toolkit”. Στο: *BMC Bioinformatics* 7.1 (Μαρ. 2006). DOI: 10.1186/1471-2105-7-170.

[4] 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.

[5] 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.

[6] 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.