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

1 ΕΙΣΑΓΩΓΗ

- Η Βιοπληροφορική έχει αναδειχθεί ως το **κομβικό επιστημονικό πεδίο** ανάμεσα στη βιολογία και την επιστήμη των υπολογιστών.
- 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 παράγονταν.
- Χρησιμοποιεί υπολογιστικά εργαλεία για την μελέτη και τη κατανόηση βιολογικών δεδομένων όπως το DNA και οι πρωτεΐνες.
- Έχει εξελιχθεί σε ένα αναγκαίο εργαλείο για ερευνητικούς σκοπούς, αφού συντελεί στην ανακάλυψη νέων φαρμάκων, στην εξατομικευμένη και προληπτική ιατρική, στη γονιδιακή θεραπεία, στη βελτιώση της καλλιέργειας κ.α., οδηγώντας σε συμπεράσματα πολύ πιο αποτελεσματικά και με μεγαλύτερη ακρίβεια.

1 ΕΙΣΑΓΩΓΗ

- Μια μεγάλη πρόκληση στη Βιοπληροφορική είναι η ανάγκη για **τυποποίηση** των τρόπων αναπαράστασης και ανταλλαγής πολύπλοκων βιολογικών δεδομένων.
- Αυτή η τυποποίηση μπορεί να λυθεί αποτελεσματικά με τη χρήση της **XML** (eXtensive Markup Language) γλώσσας.
- Πρόκειται για μια γλώσσα σήμανσης αρκετά ευέλικτη και ισχυρή για την **αναπαράσταση ιεραρχικών σχέσεων** (hierarchical relationships), κάτι αρκετά κοινότυπο στη μελέτη βιολογικών δεδομένων.
- Στη συγκεκριμένη εργασία θα δοθεί βάση στους **τρόπους αναπαράστασης βιολογικής πληροφορίας με τη χρήση XML-based γλωσσών**, τα συστήματα που γίνεται αποθήκευση και μεταφορά αυτής της πληροφορίας, ο τρόπος που είναι δομημένη, πώς μπορεί να επιτευχθεί ομογένεια λόγω των διαφορετικών πηγών πληροφορίας που οδηγεί σε αρχεία διαφορετικού τύπου μορφολογίας, κ.α.

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 προσεγγίσεις για την αναπαράσταση, ενσωμάτωση και διαχείριση βιολογικών δεδομένων με τη χρήση γλωσσών που βασίζονται στην XML. Επιπλέον, παρουσιάζεται μια νέα προσέγγιση για τη διαχείριση ετερογενών βιολογικών δεδομένων μέσω της XML.

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 αυτών, κάτι βολικό μιας και καθημερινά δημιουργούνται και αναπτύσσονται 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 επιτρέπει την αναπαράσταση με σαφήνεια πολύπλοκων βιολογικών σχέσεων.

2.1.3 Αναπαράσταση βιολογικών δεδομένων με τη χρήση XML

Για την αναπαράσταση βιολογικών δεδομένων έχουν χρησιμοποιηθεί αρκετές γλώσσες οι οποίες βασίζονται στην XML:

- **Bioinformatic Sequence Markup Language** (BSML)
- **Protein Markup Language** (ProXML)
- **RNA Markup Language** (RNAML)
- **System Biology Markup Language** (SBML)
- **Cell Markup Language** (CellML)

2.1.3.1 Bioinformatic Sequence Markup Language (BSML)

- Σχεδιάστηκε το 1997 με σκοπό να περιγράφει αλληλουχίες όπως DNA, RNA ή πρωτεϊνικές.
- Ένα BSML αρχείο περιλαμβάνει πληροφορίες για το πώς τα γονιδιώματα κωδικοποιούνται, ανακτώνονται και εμφανίζονται, ο ορισμός **χαρακτηριστικών** πάνω τους (ρυθμιστικές περιοχές, μεταλλάξεις κτλ), ενώ επιτρέπεται η **αναπαράσταση** επιπρόσθετων πληροφοριών όπως σχόλια, βιβλιογραφικές αναφορές κ.α.

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.

Biopolymer Markup Language (BioML)

```
<bioml>
  <organism>
    <species>Homo sapiens</species>
    <cell>
      <organelle type="RER">
        <particle type="ribosome">
          <protein id="1">
          </protein>
          <protein id="2">
          </protein>
          ...
          <rna id="1">
          </rna>
          <rna id="2">
          </rna>
          ...
        </particle>
      </organelle>
    </cell>
  </organism>
</bioml>
```

2.1.3.3 ΠΡΟΕΚΤΑΣΗ: RNA Markup Language - RNAML

- Γλώσσα που σχεδιάστηκε με σκοπό να διευκολύνει την ανταλλαγή RNA πληροφοριών μεταξύ διαφορετικών λογισμικών βιοπληροφορικής.
- 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 τυποποίησης της RNA πληροφορίας, με σκοπό την αύξηση της αποτελεσματικότητας στην κοινότητα των βιολόγων.
- Οι προηγούμενες προσπάθειες για τυποποίηση της βιολογικής πληροφορίας περιλαμβάνουν τη γλώσσα σήμανσης **BIOpolymer Markup Language (BIOML)** η οποία αναπτύχθηκε το 1999 από την ProteoMaterics. Περιλαμβάνει ένα 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, γενετικές τροποποιήσεις κτλ), και **δεν κάλυπτε επαρκώς πληροφορίες για τη δομή του RNA**, κάτι ερευνητικά κρίσιμο.
- Γενικότερα οι προηγούμενες προσπάθειες δεν κάλυπταν τις απαιτήσεις που έθετε η επιστημονική κοινότητα που μελετούσε το RNA, οδηγώντας στην ανάπτυξη της RNAML.

2.1.3.3 ΠΡΟΕΚΤΑΣΗ: RNA Markup Language – RNAML

- 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, μπορεί να **αναπαριστά 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** (annotation) και documentation σε κάθε στοιχείο, είναι δυνατή η **ομαδοποίηση των εμφανίσεων** του ίδιου λειτουργικού RNA σε διαφορετικούς οργανισμούς, κάνοντας εφικτή 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.
- 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), η οποία αναπαριστά ποσότητες με συμβολικά ονόματα τοπικά η καθολικά, οι **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.

2.1.3.4 ΠΡΟΕΚΤΑΣΗ: System Biology Markup Language (SBML)

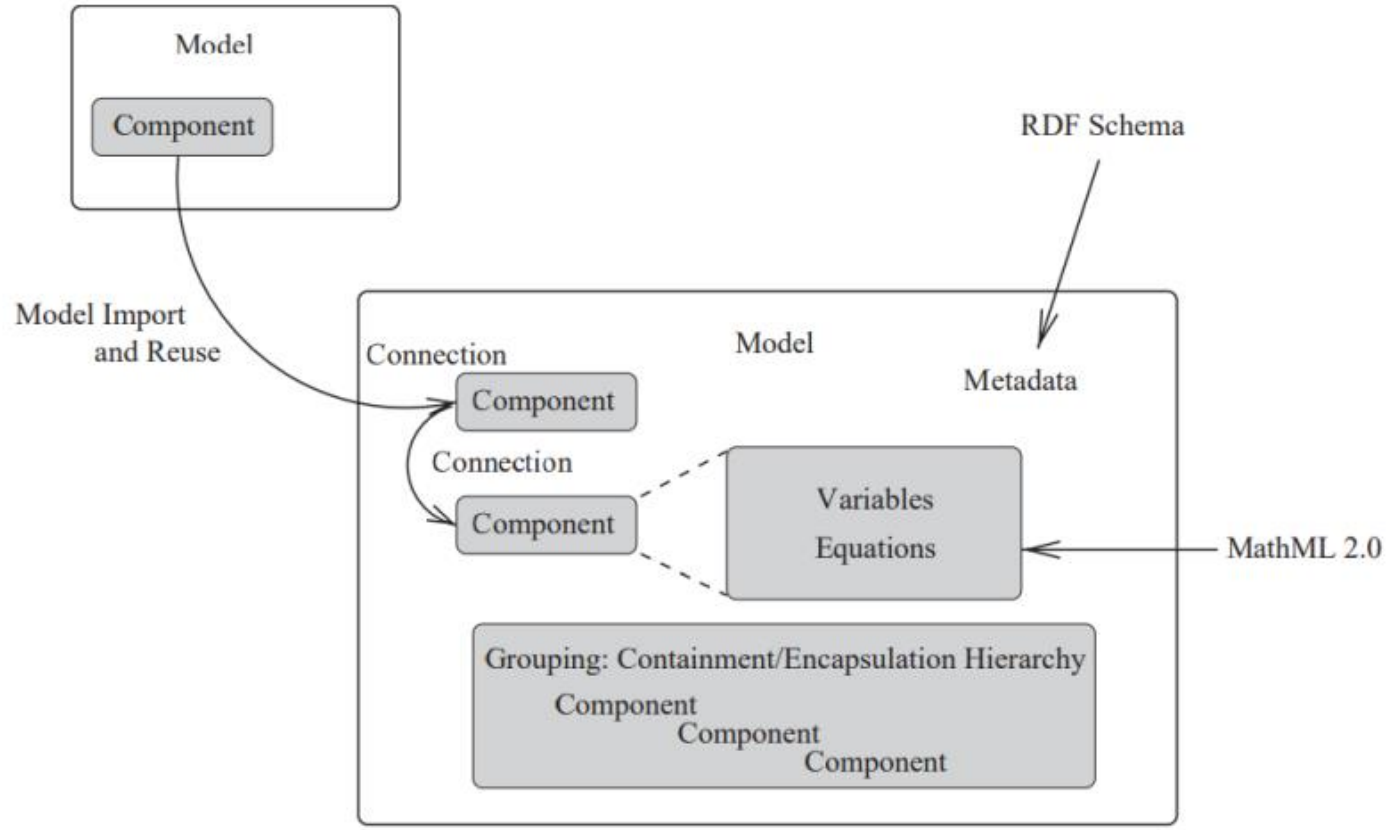
```
<?xml version="1.0" encoding="UTF-8"?>
<sbml xmlns="http://www.sbml.org/sbml/level1"
      level="1" version="2">
  <model name="gene_network_model">
    <listOfUnitDefinitions>
        ...
    </listOfUnitDefinitions>
    <listOfCompartments>
        ...
    </listOfCompartments>
    <listOfSpecies>
        ...
    </listOfSpecies>
    <listOfParameters>
        ...
    </listOfParameters>
    <listOfRules>
        ...
    </listOfRules>
    <listOfReactions>
        ...
    </listOfReactions>
  </model>
</sbml>
```

Σχήμα 2.1: Σκελετός από τον ορισμό ενός μοντέλου SBML [4]

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. 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.

2.1.3.5 ΠΡΟΕΚΤΑΣΗ: Cell Markup Language (CellML)



Σχήμα 2.2: Διάγραμμα με το σκελετό ενός CellML μοντέλου [8]

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 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.
- Ακολουθούν παραδείγματα τέτοιων συστημάτων.

2.1.4.1 DWARF

- Πρόκειται για ένα Data Warehouse σύστημα που σχεδιάστηκε για την ανάλυση μεγάλων πρωτεϊνικών οικογενειών.
- Ενσωματώνει δεδομένα που αφορούν την αλληλουχία, τη δομή και τον χαρακτηρισμό πρωτεϊνών, συνδυάζοντας δεδομένα από διαφορετικές δημόσιες βάσεις δεδομένων όπως GenBank, ExPDB, κ.α.
- 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 **πρωτεΐνη** (περιγράφει τη βιοχημική λειτουργία, τον οργανισμό προέλευσης και την ταξινόμηση των πρωτεϊνών), την **αλληλουχία των πρωτεϊνών** (σχολιασμός συγκεκριμένων θέσεων, λεπτομέρειες για μεταλλάξεις) και τη **δομή της πρωτεΐνης** (δεδομένα που σχετίζονται με τις δευτερογενείς και τριτογενείς δομές της πρωτεΐνης).

2.1.4.1 BioWarehouse

- Πρόκειται για ένα toolkit ανοιχτού κώδικα που έχει σχεδιαστεί για τη διευκόλυνση της διασύνδεσης διαφορετικών βάσεων δεδομένων βιοπληροφορικής.
- Χρησιμοποιεί τη MySQL και την Oracle ως relational database managers, και επιτρέπει την ομαλή σύνδεση διαφορετικών βάσεων δεδομένων με σκοπό να γίνονται αποτελεσματικά queries και για την εξόρυξη δεδομένων.
- 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

2.1.4.1 Atlas

- Data warehouse σύστημα που αποθηκεύει και ενοποιεί διαφορετικούς τύπους βιολογικών δεδομένων.
- Χρησιμοποιεί την SQL η οποία καλεί (μέσω API) εφαρμογές σε C++, Java και Perl γλώσσες, οι οποίες διαβάζουν πληροφορίες από άλλες βάσεις δεδομένων (GenBank, RefSeq, UniProt κ.α.) στη βάση δεδομένων του Atlas.
- Επίσης, περιλαμβάνει κάποια εργαλεία που χρησιμοποιούνται για την εξόρυξη δεδομένων.

2.1.4.1 Biozone

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

2.1.4.1 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.

2.1.4.1 ΠΡΟΕΚΤΑΣΗ: Data warehouse συστήματα

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 [9]

2.1.4.2 ΠΡΟΕΚΤΑΣΗ: Mediator-based συστήματα

- Σε αυτή την αρχιτεκτονική, οι **ξένες βάσεις δεδομένων διατηρούν την αυτονομία τους** και τα mediator-based συστήματα δρουν ως **μεσάζοντες**. Ο στόχος είναι η δημιουργία μιας ενοποιημένης προβολής των δεδομένων (globalview) **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 το **χειροκίνητο configuration του wrapper** που απαιτείται για την ενσωμάτωση των δεδομένων, αν και έχουν προταθεί κάποιες τεχνικές αυτοματοποίησης.
- Παραδείγματα τέτοιων συστημάτων είναι τα εξής:

2.1.4.2 Ontofusion

- Σύστημα οντολογίας που βασίζεται σε δύο διεργασίες: τη **χαρτογράφηση** (mapping) και την **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 UML οντολογία), η bottom-up (χτίζοντας μια νέα οντολογία) και ο συνδυασμός αυτών των δύο.
- Οι οντολογίες αυτές συνενώνονται σε ένα ξεχωριστό «global schema» όπου πλέον είναι ομογενοποιημένες.

2.1.4.2 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 βιολογικές έννοιες και ένα μοντέλο που είναι συνδεδεμένο με εξωτερικές βάσεις δεδομένων.

2.1.4.2 ΠΡΟΕΚΤΑΣΗ: Mediator-based συστήματα

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 [9]

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, 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, γίνεται σαφές πως 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.
- 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**.

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.

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.

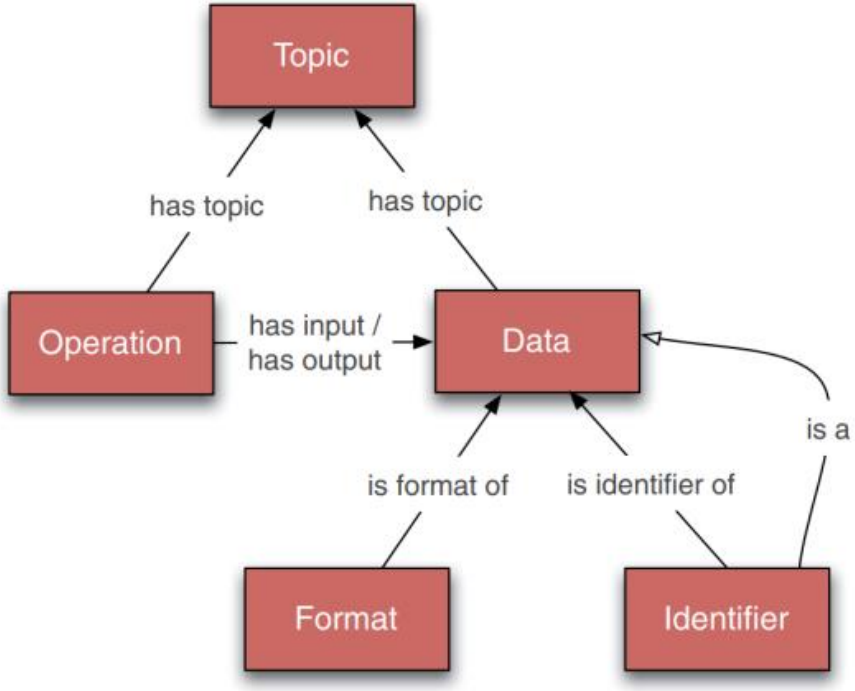
2.2.3 Τι προσφέρει η οντολογία EDAM;

- 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>`.

2.2.4 Υλοποίηση EDAM



Σχήμα 2.5: Σχέσεις μεταξύ των υποοντολογιών [6]

2.2.5 Υλοποίηση EDAM

- Η οντολογία EDAM χρησιμοποιεί δύο προσεγγίσεις για σημασιολογικό σχολιασμό (semantic annotation), τη διαδικασία κατά την οποία προσθέτουμε metadata στην πληροφορία με σκοπό την καλύτερη οργάνωση, αναζήτηση και ενσωμάτωση αυτής στις εφαρμογές μας:

2.2.5.1 Τυποποίηση σχολιασμού

- Τα XML Schemas ή τα RDF έγγραφα περιέχουν συγκεκριμένο χώρο (περιγραφές, descriptions), στο οποίο μπορούν να προστεθούν τα σχόλια.
- Κατ' αυτό τον τρόπο, τα σχόλια διαχειρίζονται από τα ίδια τα εργαλεία, είναι ανεξάρτητα, και δε χρειάζεται να επαναεγγραφούν όταν τα δεδομένα εισάγονται σε διαφορετικά frameworks, έχοντας ως αποτέλεσμα μεγαλύτερη συνοχή και αποτελεσματικότητα.

2.2.5.2 Ξεχωριστοί κατάλογοι

- Εναλλακτικά, τα σχόλια μπορούν να αποθηκευτούν σε ξεχωριστούς καταλόγους, κάτι που επιτρέπει ευελιξία στη διαχείρισή τους.

2.2.6 Χρήση EDAM σε εφαρμογές

- Η οντολογία EDAM έχει αξιοποιηθεί σε διάφορα εργαλεία βιοπληροφορικής με σκοπό την αύξηση της λειτουργικότητάς τους.
- Παραδείγματα είναι το EMBOSS Suite (όπου χρησιμοποιήθηκε για να ενισχύσει τις αναζητήσεις/queries), το eSysbio (όπου χρησιμοποιήθηκε για τη διαχείριση των δεδομένων των χρηστών με σκοπό την οπτικοποίησή τους) και το Bio-jETI (που διευκόλυνε την αυτόματη σύνθεση πληροφορίας βάσει προδιαγραφών).

2.3 METATRON

- Το άρθρο ”**MetaTron: advancing biomedical annotation empowering relation annotation and collaboration**” των Irrera, Marchesin κ.α. πραγματεύεται το εργαλείο MetaTron.
- Πρόκειται για μια εφαρμογή ανοιχτού κώδικα με σκοπό τη βελτίωση της αποτελεσματικότητας του σχολιασμού (annotation) βιοϊατρικών δεδομένων.

2.3.1 Τι είναι ο σχολιασμός (annotating);

- Πρόκειται για την **προσθήκη δομημένης πληροφορίας** σε βιοϊατρικά κείμενα με σκοπό τη **βελτίωσης της χρηστικότητάς** τους για ερευνητικούς σκοπούς ή για κλινικές εφαρμογές.
- Συνήθως περιλαμβάνει τον **εντοπισμό και την επισήμανση γονιδίων**, πρωτεϊνών και άλλων οντοτήτων, τον **καθορισμό σχέσεων** μεταξύ τους (πώς ένα φάρμακο αλληλεπιδρά με μια ασθένεια ή πως σχετίζονται δύο γονίδια μεταξύ τους).
- Για να επιτευχθεί, χρησιμοποιούνται οντολογίες όπου βοηθούν στον σαφή καθορισμό των σχέσεων των εννοιών, μετατρέποντας τις έννοιες σε machine-readable, οδηγώντας στον αυτοματισμό και στην εξόρυξη γνώσης.

2.3.1.1 ΠΡΟΕΚΤΑΣΗ: Κριτήρια (χειροκίνητου) σχολιασμού

- Ο **χειροκίνητος σχολιασμός**, δηλαδή ο σχολιασμός που προέρχεται από τους ίδιους τους επιστήμονες είναι μια **διαδικασία κουραστική και χρονοβόρα**.
- Απαιτεί εξαιρετική εξειδίκευση από τους επιστήμονες ώστε να μπορέσουν να ταξινομήσουν με ακρίβεια τις οντότητες, να ελέγξουν για λάθη, κάτι το οποίο κοστίζει όλο και περισσότερο όσο αυξάνεται η πολυπλοκότητα και το μέγεθος των δεδομένων.
- Υπάρχουν πάρα πολλά κριτήρια που επιζητούνται από τα λογισμικά που χρησιμοποιούνται για χειροκίνητο σχολιασμό, που αφορούν τα τεχνικά χαρακτηριστικά τους, τη χρηστικότητά τους, κ.α.
- Παραδείγματα για τα τεχνικά χαρακτηριστικά είναι η διαθεσιμότητα του κώδικα, η ευκολία εγκατάστασης, η ποιότητα του documentation, το κόστος, ενώ για τη χρηστικότητά τους οι σημειώσεις πολλαπλών ετικετών (multi-label annotations), ενσωμάτωση με οντολογίες, προσχολιασμούς βάση προηγούμενων δεδομένων, απόρρητο δεδομένων κα.

2.3.1.1 ΠΡΟΕΚΤΑΣΗ: Κριτήρια (χειροκίνητου) σχολιασμού

Έρευνα που έγινε για τα σημαντικότερα χαρακτηριστικά ενός λογισμικού σχολιασμού, κατέληξε ότι τα σημαντικότερα είναι:

- να είναι διαθέσιμο διαδικτυακά ή να μπορεί να εγκατασταθεί εύκολα
- να είναι λειτουργικό, με διαισθητικές λειτουργίες χωρίς να απαιτεί μεγάλο επίπεδο εμπειρίας από τον χρήστη
- να υποστηρίζει σχηματική αναπαράσταση με ευέλικτα εργαλεία που να καλύπτουν πολλά διαφορετικά use-cases

Τα λογισμικά σχολιασμού μέχρι τώρα (και αυτά που βασίζονται στη βιοϊατρική, και γενικού σκοπού) δεν κάλυπταν όλα αυτά τα κριτήρια και χαρακτηριστικά ταυτόχρονα, με κάποια να υπερτερούν σε μερικά και να υστερούν σε άλλα. Επομένως, είναι σαφής η ανάγκη για έναν πιο αποτελεσματικό, λογισμικό σχολιασμού.

2.3.2 Ο ρόλος του Metatron

- Το Metatron είναι από τα λίγα εργαλεία σχολιασμού που καταφέρνει να είναι αποτελεσματικό σε όλα τα προαναφερόμενα χαρακτηριστικά.
- Υποστηρίζει διαφορετικούς τύπους αρχείων, μπορεί να συνδεθεί με APIs από πήγες όπως PubMed για πρόσβαση σε επιπλέον πηγές, περιλαμβάνει διαφορετικούς τύπους σχολιασμού, δυνατότητες για συνεργατικό σχολιασμό, αυτόματες προτάσεις, παραμετροποίηση κα.

2.3.3 Χαρακτηριστικά του MetaTron

- Το MetaTron υποστηρίζονται **πολλαπλοί τύποι σχολιασμού**, σχολιασμός σε **επίπεδο εγγράφου** (documentlevel) που περιλαμβάνουν την ανάθεση ετικετών σε ολόκληρα έγγραφα, και σε **επίπεδο αναφοράς** (mentionlevel) που εστιάζουν σε συγκεκριμένα τμήματα του κειμένου.
- Ο σχολιασμός σε επίπεδο κειμένου περιλαμβάνει σχόλια-ετικέτες (labels) και σχόλια-ισχυρισμούς (assertions), τα οποία μπορούν να συμπεριληφθούν σε RDF γραφήματα για καλύτερη αναπαράσταση.
- Υποστηρίζονται οι οντολογίες, επιτρέποντας στους χρήστες να ορίζουν **εννοιών** (concepts), ο συνεργατικός σχολιασμός, **πολλαπλοί τύποι αρχείων** και μεγάλη παραμετροποίηση. Επιπλέον, περιλαμβάνει το AutoTron, ένα χαρακτηριστικό που προσφέρει προβλέψεις στο σύστημα για τον **αυτοματοποιημένο σχολιασμό**, σκοπεύοντας στην ενίσχυση της αποτελεσματικότητας των χρηστών.

2.3.3.1 Αρχιτεκτονική του MegaTron

- Η αρχιτεκτονική του MegaTron χωρίζεται σε τρία επίπεδα, το επίπεδο δεδομένων (data layer), το επιχειρησιακό επίπεδο (business layer) και το επίπεδο παρουσίασης (presentation layer).
- Το επιχειρησιακό επίπεδο χρησιμοποιεί ένα REST API σε Django Python, δρώντας ως ο μεσολαβητής ανάμεσα στο επίπεδο παρουσίασης και επίπεδο δεδομένων. Το επίπεδο παρουσίασης αναπτύχθηκε χρησιμοποιώντας ReactJS, HTML/CSS/JS.

2.3.4 Υλοποίηση και αποτελέσματα

- Το άρθρο μπαίνει σε μια λεπτομερή περιγραφή των χαρακτηριστικών και του τρόπου λειτουργίας του λογισμικού, όπως επίσης και έρευνα των χρηστών για το πόσο έμειναν ικανοποιημένοι, πράγματα που ανήκουν εκτός της σφαίρας της μελέτης μας.

2.4 ΣΥΝΑΛΛΑΓΕΣ ΣΕ ΣΥΣΤΗΜΑΤΑ ΔΕΔΟΜΕΝΩΝ ΜΕΓΑΛΗΣ ΚΛΙΜΑΚΑΣ

- Η συλλογή με τίτλο "**Transactions on Large-Scale Data and Knowledge Centered Systems XXIV**" περιλαμβάνει ένα σύνολο από δημοσιεύσεις:

2.4.0.1 Reflective Constraint Writing: A Symbolic Viewpoint of Modeling Languages

- Παρουσιάζεται ένας τρόπος για την επέκταση των object constraint γλωσσών (περιγράφουν κανόνες που ισχύουν σε UML μοντέλα) μέσω του reflection, επιτρέποντας έτσι τα metadata να είναι διαθέσιμα στο επίπεδο του αντικειμένου.

2.4.0.2 PPP-Codes for Large-Scale Similarity Searching

- Παρουσιάζεται ένας τρόπος για να αντιμετωπιστεί η δυσκολία του αποτελεσματικού εντοπισμού παρόμοιων αντικειμένων σε μεγάλους χώρους αναζήτησης.
- Επιτυγχάνεται μέσω μιας αναζήτησης δύο φάσεων που χρησιμοποιεί μια νέα δομή δεδομένων που ονομάζεται δείκτης PPP-Code, ο οποίος υπολογίζεται ανεξάρτητες κατατάξεις (rankings) χρησιμοποιώντας μια συνάρτηση απόστασης και συγκεντρώνει αυτές τις κατατάξεις αυξάνοντας την αποτελεσματικότητα της αναζήτησης.

2.4.0.3 Solving Data Mismatches in Bioinformatics Workflows by Generating Data Converters

- Λόγω της ετερογένειας των δεδομένων στη βιοπληροφορική, συχνά παρουσιάζονται αναντιστοιχίες
- (mismatches) μεταξύ των εισόδων και των εξόδων σε διαφορετικά λογισμικά και υπηρεσίες, κάτι που καθιστά δύσκολη την δουλειά των επιστημόνων. Μέχρι πρότινος, ένας τρόπος για την ενοποίηση των δεδομένων είναι οι slims μετατροπείς, που είναι χρονοβόρο να τους γράψεις με το χέρι, και όταν δημιουργούνται αυτόματα δεν είναι αποτελεσματικοί.
- Το άρθρο παρουσιάζει ένα νέο τρόπο για τη συστηματική μετατροπή των εξόδων σε εισόδους, χρησιμοποιώντας ένα σύστημα κανόνων παρόμοιο με το XML Schema.

2.4.0.4 A Framework for Sampling-Based XML Data Pricing

- Παρουσιάζεται ένα framework για την τιμολόγηση XML εγγράφων, θέτοντας ένα βάρος σε κάθε έγγραφο.

2.4.0.5 kdANN+: A Rapid AkNN Classifier for Big Data

- Παρουσιάζεται ο kdANN+, ένας αποδοτικός ταξινομητής K-Nearest-Neighbour, σχεδιασμένος για big data εφαρμογές. Παρουσιάζονται αλγόριθμοι που βελτιώνουν την ταχύτητα και την ακρίβεια της ταξινόμησης (classification) σε μεγάλα σύνολα δεδομένων, κάνοντάς τον κατάλληλο για real-time εφαρμογές.

2.4.0.6 Optimizing Inter-data-center Large-Scale Database Parallel Replication with Workload-Driven Partitioning

- Εισάγεται μια partitioning στρατηγική με σκοπό τη βελτιστοποίηση παράλληλης αντιγραφής σε distributed big-data βάσεις δεδομένων.

2.4.0.7 Anonymization of Data Sets with NULL Values

- Διερευνώνται μεθόδους για την ανωνυμοποίηση συνόλων δεδομένων που περιλαμβάνουν NULL τιμές.

2.4.1 Αυτόματος μετατροπέας

- Παρουσιάζεται ένας αυτόματος μετατροπέας που βασίζεται σε έναν μηχανισμό κανόνων που ανιχνεύει αν υπάρχει μετατρεψιμότητα μεταξύ διαφορετικών τύπων δεδομένων.
- Ορίζεται ένα σύνολο τύπων αναπαράστασης που βασίζονται σε type constructors, όπου ο καθένας τους θέτει ένα σύνολο από XML τιμές.
- Παραδείγματα κανόνων μετατρεψιμότητας είναι: `PRIMITIVE`, `TAGCHANGE`, `TAGREMOVAL`, `EMPTY`, `CONCAT`, `LEFTSELECTION`, `MAP` και άλλοι. Αυτοί δημιουργούν αντίστοιχες συναρτήσεις που μετατρέπουν τις εισόδους σε XML.

2.4.1 Αυτόματος μετατροπέας

Function	Input	Output	Description
Content	XML	XML	Returns the content of an XML element
Element	Tag, XML	XML	Builds an XML element given a tag and an XML content
Concat	XML, XML	XML	Returns the concatenation of two XML sequences
Select	XML, type, type	XML, XML	Splits an XML sequence in two parts matching given types
Map	Converter, XML	XML	Applies a converter to each node of an XML sequence and returns the concatenation of the results
Choose	XML, type	XML	Returns any element matching a given type from an XML sequence

Σχήμα 2.6: Σύνολο συναρτήσεων [3]

2.4.1 Αυτόματος μετατροπέας

- Όταν εντοπιστεί η σχέση μεταξύ των τύπων δεδομένων, το σύστημα δημιουργεί αυτόματα τους κατάλληλους μετατροπείς. Οι μετατροπείς ελέγχονται για τη βιολογική εγκυρότητά τους, και εν τέλει ενσωματώνονται στο workflow των ερευνητών, επιτρέποντας την απρόσκοπτη ροή δεδομένων μεταξύ διαφορετικών υπηρεσιών.

Βιβλιογραφία

1. EDAM - Ontology of bioscientific data analysis and data management. URL: https://edamontology. org/.
2. 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.
3. Abdelkader Hameurlain κ.ά. Transactions on large-scale data- and knowledge-centered systems XXIV: Special issue on database- and expert-systems applications. Springer, 2016.
4. 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.
5. Ornella Irrera, Stefano Marchesin και Gianmaria Silvello. “Metatron: Advancing Biomedical Annotation Empowering Relation Annotation and collaboration”. Στο: BMC Bioinformatics 25.1 (Μαρ. 2024). DOI: 10.1186/s12859-024-05730-9.
6. Jon Ison κ.ά. “Edam: An ontology of bioinformatics operations, types of data and identifiers, topics and formats”. Στο: Bioinformatics 29.10 (Μαρ. 2013), σσ. 1325–1332. DOI: 10 . 1093 / bioinformatics / btt113.
7. 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.
8. 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.
9. 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.
10. Mariana Neves και Jurica eva. Στο: Briefings in Bioinformatics 22.1 (Δεκ. 2019), σσ. 146–163. DOI: 10.1093/bib/bbz130.
11. Ontology (information science). Αύγ. 2024. URL: https://en.wikipedia.org/wiki/Ontology_ (information_science)
12. D. Pérez-Rey κ.ά. “Ontofusion: Ontology-based integration of genomic and clinical databases”. Στο: Computers in Biology and Medicine 36.7–8 (Ιούλ. 2006), σσ. 712–730. DOI: 10.1016/j.compbiomed.2005.02.004
13. Nadine Schuurman και Agnieszka Leszczynski. “Ontologies for Bioinformatics”. Στο: Bioinformatics and Biology Insights 2 (Ιαν. 2008). DOI: 10.4137/bbi.s451
14. Sohrab P Shah κ.ά. “Atlas – A Data Warehouse for Integrative Bioinformatics”. Στο: BMC Bioinformatics 6.1 (Φεβ. 2005). DOI: 10.1186/1471-2105-6-34.
15. Robert Stevens κ.ά. “Tambis: Transparent access to multiple bioinformatics information sources”. Στο: Bioinformatics 16.2 (Φεβ. 2000), σσ. 184–186. DOI: 10.1093/bioinformatics/16.2.184
16. 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