JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 24, Number 10, 2017
Mary Ann Liebert, Inc.
Pp. 969–980
DOI: 10.1089/cmb.2016.0095

EpiGeNet: A Graph Database of Interdependencies Between Genetic and Epigenetic Events in Colorectal Cancer

IRINA BALAUR,[1] MANSOOR SAQI,[1] ANA BARAT,[2] ARTEM LYSENKO,[3] ALEXANDER MAZEIN,[1] CHRISTOPHER J. RAWLINGS,[3] HEATHER J. RUSKIN,[4] and CHARLES AUFFRAY[1]

ABSTRACT

The development of colorectal cancer (CRC)—the third most common cancer type—has been associated with deregulations of cellular mechanisms stimulated by both genetic and epigenetic events. StatEpigen is a manually curated and annotated database, containing information on interdependencies between genetic and epigenetic signals, and specialized currently for CRC research. Although StatEpigen provides a well-developed graphical user interface for information retrieval, advanced queries involving associations between multiple concepts can benefit from more detailed graph representation of the integrated data. This can be achieved by using a graph database (NoSQL) approach. Data were extracted from StatEpigen and imported to our newly developed EpiGeNet, a graph database for storage and querying of conditional relationships between molecular (genetic and epigenetic) events observed at different stages of colorectal oncogenesis. We illustrate the enhanced capability of EpiGeNet for exploration of different queries related to colorectal tumor progression; specifically, we demonstrate the query process for (i) stage-specific molecular events, (ii) most frequently observed genetic and epigenetic interdependencies in colon adenoma, and (iii) paths connecting key genes reported in CRC and associated events. The EpiGeNet framework offers improved capability for management and visualization of data on molecular events specific to CRC initiation and progression.

Keywords: computational molecular biology, graph database, epigenetics, molecular interdependencies, colorectal cancer, networks.

[1]European Institute for Systems Biology and Medicine (EISBM), CIRI UMR CNRS 5308, CNRS-ENS-UCBL-INSERM, Université Claude Bernard, Lyon, France.
[2]Department of Physiology and Medical Physics, Centre for Systems Medicine, Royal College of Surgeons in Ireland, Dublin, Ireland.
[3]Rothamsted Research, Hertfordshire, United Kingdom.
[4]Centre for Scientific Computing and Complex Systems Modelling, School of Computing, Dublin City University, Dublin, Ireland.

1. INTRODUCTION

ACCORDING TO STATISTICAL STUDIES, colorectal cancer (CRC) is the third most common cancer type, after lung and breast cancers, accounting for around 8% and 13% of all new cancer cases in the United States (2015) (Cancer of the Colon and Rectum—SEER Stat Fact Sheets, n.d.) and Europe (2012) (Bowel Cancer Incidence Statistics, n.d.), respectively. Over the last two decades, it has been shown that both genetic and epigenetic events (e.g., mutation, deletion, insertion of DNA sequences, and molecular interactions that affect gene expression without changing DNA sequence) induce abnormal micromolecular modifications, leading to cancer development (further details are given in the Brief Biological Background section).

A major characteristic of cancers, which affects both investigation and interpretation, is heterogeneity of the biological information that characterizes malignant systems. The representation of these systems involves relationships linking multiple interdependencies between genetic and epigenetic modifications that lead to cancer development. Relational databases that focus on various features of cancer pathways have appeared with increasing frequency [e.g., MethyCancer (He et al., 2008), Catalogue of Somatic Mutations in Cancer (COSMIC) (Forbes et al., 2015)]. While the relational data approach has proved useful for management of structured data, issues remain in moving forward. For example, (i) integration of multiple data types in relational databases is nontrivial as it involves redefinition of data schema where new information follows a different structure; (ii) in exploration of cancer-related hypotheses, different concepts may need to be linked. Using highly connected information (specific to malignant systems) can be inefficient for queries, which associate (join) data stored in a number of different tables with correspondingly slow response times. The graph database approach has demonstrated capability in facilitating not only both (i) and (ii) but also in (iii) exploration of context, characterized by wide diversity.

In a graph database, concepts are represented by nodes and their associations by edges. Thus, the approach provides a more natural way of representing highly interconnected data, with exploration of stored content benefitting additionally from the use of different graph algorithms. A key feature of graph databases is that traversal exploration is permitted [i.e., nodes can be accessed from neighbor nodes by means of edge connections (relationships)]. This gives a major advantage in terms of performance in comparison to relational databases, where such associations require loops on multiple table indexes. Examples of graph database frameworks include AllegroGraph (AllegroGraph url, n.d.), Sparksee (Sparksee url, n.d.), FlockDB (FlockDB url, n.d.), InfiniteGraph (InfiniteGraph url, n.d.), OrientDB (OrientDB url, n.d.), and Neo4j (Neo4j url, n.d.).

Neo4j is a well-established framework for ''property graphs,'' fundamental to graph databases specialized for directional relationships (directed edges) and ''multirelational graphs'' (with nodes linked by multiple different edges). Additional information on concept and relationships can be stored as properties (or attributes) of the nodes and edges. The Neo4j framework uses Cypher, a declarative query language (similar to SQL), to perform data interrogation. In Life Sciences, the Neo4j framework has been used to develop ecosystems that facilitate management (integration, visualization, and exploration) of various biological and medical data types. For example, HitWalker2 is an interactive framework that integrates different data types (such as gene expression, DNA methylation, and drug sensitivity) and can be used to investigate gene context in human diseases (Bottomly et al., 2015). HRGRN is a Neo4j-based framework developed for management of genome-scale data related to Arabidopsis systems (including information on metabolic and signaling pathways, gene regulation), which facilitates investigation of relationships (associations, interactions) among these data (HRGRN url, n.d.).

Neo4j-based models have been developed also to capture and explore semantic relationships among computational and mathematical models related to cancer and to other biological systems [e.g., Johnson et al. (2014) and Henkel et al. (2015)]. In Henkel et al. (2015), authors describe a Neo4j-based framework that facilitates identification, comparison, and ranking of in-silico models (encoded in SBML and CELLML standard formats and stored in major specialized resources such as BioModels Database) that correspond to specific categories.

In this study, we present a Neo4j graph database developed for the management of genetic–epigenetic interdependencies in CRC development. We provide Cypher query examples on the way in which the graph database can be applied to CRC initiation to identify (i) genetic–epigenetic modifications and (ii) molecular phenomena observed and reported in the specialized literature. In addition, we explore path connections associated with the highest ''incidence score''; the score is computed as a product of the conditional probabilities of relationships between molecular events, with the highest scores associated with the most plausible pathways.

2. BRIEF BIOLOGICAL BACKGROUND

CRC initiation is associated with aberrant cell growth rate in the colon epithelium leading to polyps (considered benign). If not removed, the colon adenomatous polyps may increase in size and become malignant over time. Thus, while the colon *adenoma* is a phenotype, characterized by benign modifications, malignant characteristics are already present at adenocarcinoma or carcinoma stages. As the tumor progresses, cancer cells feature the accumulative aberrant changes within polyps that facilitate cell proliferation and eventually migration. Finally, CRC can extend to other organs, including the liver and lungs, leading to metastasis.

Cancer initiation and progression have been linked, in recent years, to aberrant genetic and epigenetic changes. Epigenetic events are molecular phenomena that influence gene expression without modifying the DNA sequence [e.g., Allis et al. (2007)]. Epigenetic modifications have been observed to occur as part of the aging process and in the earliest stages of human diseases, including cancers and neurodegenerative disorders. Signatures include changes in DNA methylation, proteins known as histones (that contribute to nucleosome arrangement of the DNA sequence), and small noncoding RNAs (which contribute to cell protection) (Allis et al., 2007; Baylin and Jones, 2011). DNA methylation (DNAm) (or the addition of a methyl group to a cytosine ring) is a major epigenetic event with an important role in gene regulation (Allis et al., 2007). Two aberrant forms of DNA methylation, *hyper*- and *hypo*methylation (increased and decreased methylation relative to normal, respectively), have been detected in cancer development. Specifically, *hyper*methylation of the CpG islands[1] in the promoter of the tumor suppressor genes leads, in many cases, to gene silencing, while global hypomethylation influences proto-oncogene activation and chromosomal instability (Bjornsson et al., 2004; Allis et al., 2007; Baylin and Jones, 2011). Modification of histones is another major epigenetic event influencing chromatin dynamics (where chromatin is the combination of DNA and proteins that comprise the cell nucleus). Acetylation[2] and methylation of histones H3 and H4 (known as core[3] histones) are the most studied forms of modification to date.

Interdependency between DNAm and histone modifications also has been recently reported (Cedar and Bergman, 2009). Specifically, findings indicate that unmethylated DNA and histone acetylation determine an open chromatin form, while nonacetylated histones and DNAm induce a more compact chromatin structure. In addition, histone methylation can increase DNAm level. In terms of the dynamics, DNAm is known to change more slowly than histone proteins (Cedar and Bergman, 2009). In addition, small noncoding RNAs play a major role in cellular developmental phases, being involved in cell protection against viral infections and also in determining DNA methylation patterns (Carthew and Sontheimer, 2009; Ghildiyal and Zamore, 2009; Mattick et al., 2009).

Epigenetic modifications are notable, both for their reversibility potential and for faster dynamics compared to genetic alterations (Dworkin et al., 2009; Alegría-Torres et al., 2011). Over the last decade, development of drugs targeting different epigenetic changes has become a major area of interest for pharmaceutical companies. In epigenetic therapy, the focus is thus to identify molecular mechanisms, which can inhibit epigenetic alteration occurring or succeed in reversing that which has taken place, while minimizing side effects of dosage (Azad et al., 2013; Stein, 2014).

3. METHODS

3.1. Data collection

The EpiGeNet framework has been developed using a graph database approach by integrating statistical data on molecular interdependencies observed in CRC development, mined from a manually curated and annotated database, StatEpigen (Barat and Ruskin, 2010; StatEpigen url, n.d.) (note: reference date for data integration from StatEpigen into EpiGeNet is November 30, 2015). In StatEpigen, information is structured by simple and conditional relationships between genetic and epigenetic events. Data on hyper/hypomethylation, mutation, histone modifications, loss of heterozygosity, and gene expression are included (for healthy phenotype) with

[1]CpG islands are genomic regions (of length ≥ 200 bp) with the percentage of CG dinucleotides $> 50\%$.
[2]Histone acetylation and methylation refer to addition of acetyl and methyl compounds, respectively, to histones.
[3]Two groups of histones are known: the core set (including H2A, H2B, H3, H4) and the linker set (H1 and H5) (Ito, 2007).

additional detail on polyps, adenoma, carcinoma, or metastasis (for CRC development in aberrant cases). Note: the baseline data extracted from StatEpigen use the gene symbols (HUGO notation); full gene names used in this article are provided in Supplementary Table S1, where the gene symbol-full name mapping was resolved using the ''Retrieve/ID mapping'' tool from the UniProt database (Consortium, 2015). The simple relationships represent the probability of single molecular event occurrence at a specific oncogenesis stage; for example, from StatEpigen (Barat and Ruskin, 2010),

$$\mathrm{P}((\mathrm{APC\ hyperMeth\_CpG}), \mathrm{adenoma}) = 0.459, \tag{1}$$

is the probability of hypermethylation at CpG islands in APC promoter = 0.459 in colon adenoma for the available data set. Simple relationships have general form giving values (val1)

$$\mathrm{P}((\mathrm{Ge}), \mathrm{s}) = \mathrm{val1}, \tag{2}$$

where G = gene symbol, e = molecular event for gene G, s = oncogenesis stage. Thus, in Equation 1, G = *APC* gene, e = hypermethylation at promoter CpG islands, and s = adenoma. The conditional relationships are given by the Bayesian expression for the dependence of two molecular events, described by:

$$\mathrm{P}((\mathrm{G2\ e2})|(\mathrm{G1\ e1}), \mathrm{s}) = \mathrm{val2}, \tag{3}$$

where G1, G2 = gene symbols, e1, e2 = molecular events for genes G1 and G2, respectively, s = oncogenesis stage, as before. For example, from StatEpigen (Barat and Ruskin, 2010),

$$\mathrm{P}((\textit{KRAS}\ \mathrm{mutation})|(\textit{APC}\ \mathrm{hyperMeth\_CpG}), \mathrm{adenoma}) = 0.269, \tag{4}$$

is the conditional probability of *KRAS* mutation (based on empirical evidence from the literature), given *APC* hypermethylation at CpG islands in adenoma stage. Terms are defined similarly as above. In addition, conditional relationships between various events observed in the context for the same gene can be obtained from the curated literature and are available also in StatEpigen. For example, the relationship

$$\mathrm{P}(\textit{KRAS}\ \mathrm{mutation}|\mathrm{KRAS\ mutation}, \mathrm{polyps}) = 0.330 \tag{5}$$

indicates the empirical conditional probability (relative frequency) between two different mutation types of the *KRAS* gene, in polyps phenotype = 0.330. In this article, such conditional relationships are denoted as ''self-relationships'' (where G1 = G2 and e1 = e2). For both simple and conditional relationships, the molecular event is denoted by the pairwise list gene symbol (G) and event type (e).

3.2. Data model

In EpiGeNet, molecular events of conditional relationships are represented by the MolecularEvent nodes. The node identifier (key) is given by a pairwise (Gi ei) list; the relationship between molecular events is represented by an edge connecting the two MolecularEvent nodes. The edge type is determined by phenotype information (healthy phenotype or aberrant stages, including polyps, adenoma, carcinoma, metastasis) and edge direction by the conditionality of the relationship. Information on gene symbol and event type is stored as attributes of the MolecularEvent node, and the probability value is stored as a property of the edge (denoted as ''CondProbValue''). Details on the experiments, providing information on the conditional relationship, are also stored as edge properties (attributes); specifically, these include the ''TestedSampleNo'' attribute (number of samples of the experimental analysis) and the ''Pmid'' attribute (the PubMed identifier of the publications describing the experiments). For example, the conditional relationship shown in Equation 4 is represented by two MolecularEvents nodes, connected by an ADENOMA edge with direction *APC* hypermethylation -> *KRAS* mutation, where CondProbValue = 0.269 (Fig. 1). The relationship (Equation 4) was reported in article with PubMed identifier (Pmid) = 16902913 (Judson et al., 2006), where *APC* hypermethylation was measured in 52 tumor samples. Thus, edge attribute: Pmid = 16902913 and TestedSampleNo = 52. The ''self-relationships'' (e.g., Equation 5) are represented similarly, with the edge linked in this case to the same node (Fig. 1).

The StartEvent node label was introduced for representation of simple relationships [such as expression (Equation 1)] to facilitate distinction between these and self-relationship terms. Thus, a simple relationship is represented by an edge from a StartEvent node to a MolecularEvent node, where both nodes contain the same information on gene symbol and event type. Information on phenotype levels indicates edge

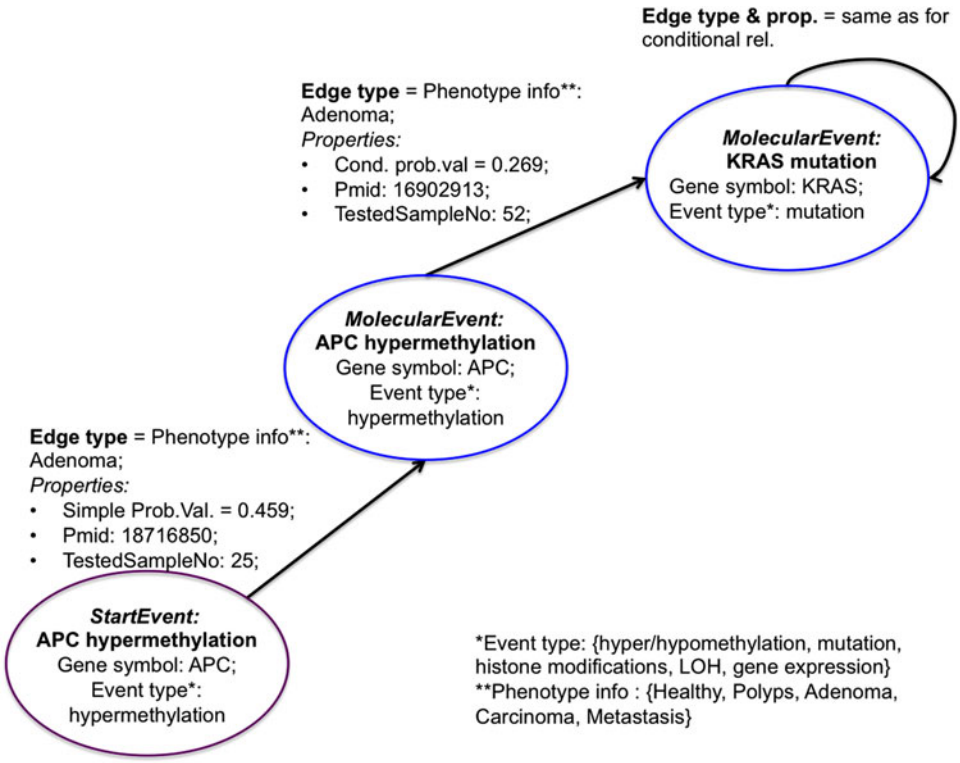

FIG. 1. The data model representation. Schematic representation of genetic–epigenetic interdependencies in healthy phenotype and aberrant phenotype, indicating different stages of colon oncogenesis. The conditional relationships are represented by edges connecting two MolecularEvents nodes (blue circles); the simple relationships are represented by edges connecting a StartEvent node (violet circle) and a MolecularEvent node (blue circle). The edge type is given by the phenotype [i.e., healthy, aberrant (polyps, adenoma, carcinoma, or metastasis)], and the edge direction is indicated by the conditionality of the event relationship. Relationship probability value is stored as an edge attribute. Information on gene name and event type is stored as node attributes. Event types are genetic or epigenetic signals, including hyper/hypomethylation, mutation, histone modifications, gene expression, and loss of heterozygosity. In addition, details on experimental analysis (including publication identifier in the PubMed database and the sample number considered) are stored as edge attributes.

type, with simple probability value stored as an attribute of the edge (denoted as the ''SimpleProbValue''). Similar to conditional relationship, details on the publication identifier and sample number of the experimental analysis are stored as ''Pmid'' and ''TestedSampleNo'' edge attributes, respectively. Hence, the simple relationship shown in Equation 1 (which indicates the probability of gene *APC* being hypermethylated when phenotype is adenoma) is represented by an ADENOMA edge connecting a StartEvent node (with key = *APC* hypermethylation) to a MolecularEvent node with the same key (Fig. 1); this edge has attributes, Pmid = 18716850 (Dhir et al., 2008) and TestedSampleNo = 25.

Data from StatEpigen were filtered to include only results where both molecular events are present in the conditional relationships and information on the tested sample number is given. In cases where different studies have reported the same conditional/simple relationships, but with different probability values, these data were combined to give a single conditional/simple relationship with probability value calculated from a weighted arithmetic mean (expectation) based on the initial probability values and tested sample numbers. The data model is represented in Figure 1, and details on the number of edges corresponding to the phenotypes are given in Table 1.

3.3. Availability of the database

The framework described above was developed predominantly in JAVA (eclipse) using the Neo4j 2.3.1 functionality. It is available for noncommercial purposes, and the code files developed to populate the Neo4j graph database using the StatEpigen data are freely available (see the first Reference). The queries described further can be explored (see first Reference for url).

Table 1. Information on Types and Occurrences of Relationships in the EpiGeNet Framework

Edge type in EpiGeNet	*Simple relationship frequency of occurrence*	*Conditional relationship frequency of occurrence*
Healthy	53	0
Polyps	40	10
Adenoma	116	45
Carcinoma	240	287

4. RESULTS AND DISCUSSION

4.1. Identification of molecular conditional relationships observed in adenoma, but not reported for polyps phenotype (according to StatEpigen data)

Identification of abnormal molecular modifications, specific to adenoma (from empirical StatEpigen sources) together with corresponding polyps phenotype, may facilitate understanding of mechanisms leading to CRC initiation. In Listing 1, we use the Cypher language to address this question and to query events that are exclusively related to adenoma with no polyps phenotype information (based on StatEpigen curation).

Listing 1. *Cypher query to identify molecular conditional relationships, observed in adenoma, but not matched with the polyps phenotype (integrated data basis—StatEpigen curation):*

```
MATCH (m1:MolecularEvent)-[r:ADENOMA]->(m2:MolecularEvent) where not (m1)-[:
POLYPS]->(m2) RETURN r
```

The results, for Listing 1, include a set of 43 conditional relationships between genetic–epigenetic events in adenoma, with multiple interdependencies between signals affecting *MGMT*, *APC*, *TP53*, *KRAS*, and *CDKN2A* (*P14* and *P16*) genes. The query can be modified to explore differences between different colorectal phenotypes

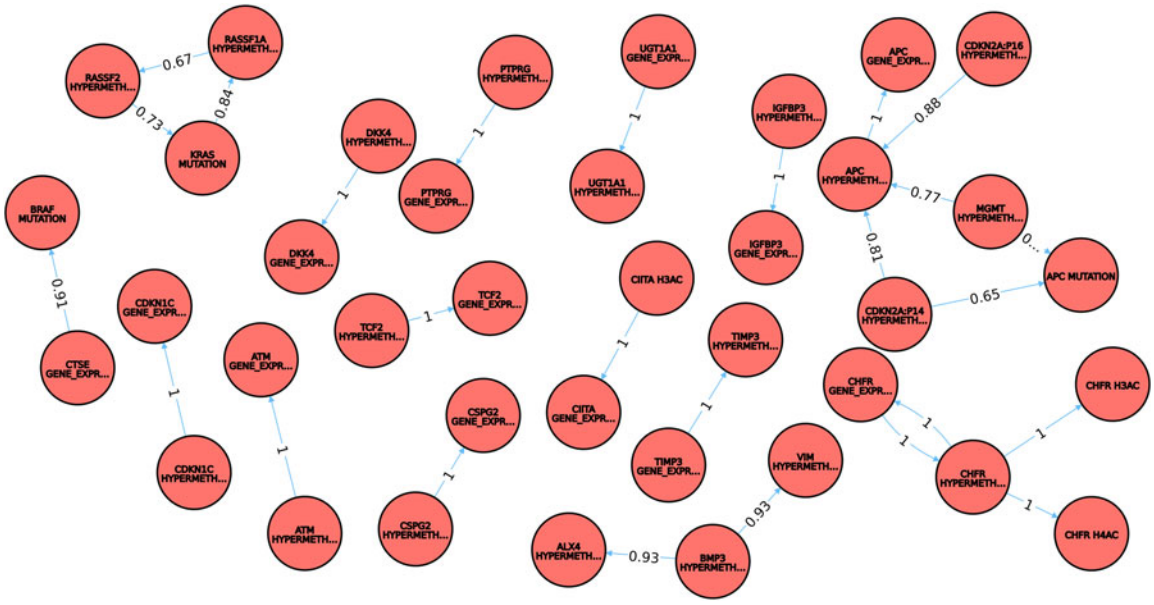

FIG. 2. Subnetwork containing molecular conditional relationships observed in adenoma but unreported for the polyps phenotype; only the conditional relationships with probability T >0.60 (Listing 2) are displayed. Legend: nodes: epigenetic and genetic events (e.g., APC hypermethylation, KRAS mutation); edges: colon adenoma (i.e., the phenotype where the molecular signals were observed). Information on probability value of the conditional relationships is stored and displayed as edge's attribute.

(e.g., adenoma and carcinoma, carcinoma and metastasis) or to include filters on probability values [e.g., all conditional relationships for which the probability is above a given threshold, T = 0.60 (Listing 2)]. With this probability threshold, results from Listing 2 include a set of 26 conditional relationships only, where the subset of the multiply-connected molecular events affects *APC*, *MGMT*, and *CDKN2A* (*P14* and *P16*) genes (Fig. 2).

Listing 2. *Cypher query to identify molecular conditional relationships, observed in adenoma, but not matched with polyps phenotype, and with probability value = T > 0.60 (based on StatEpigen curation):*

```
MATCH (m1:MolecularEvent)-[r:ADENOMA]->(m2:MolecularEvent) where not (m1)-
[:POLYPS]->(m2) and r.CondProbValue > 0.60 RETURN r
```

4.2. Identification of molecular event neighborhood (in terms of connected nodes) in colon carcinoma

Over the last two decades, several key molecular events in CRC development have been identified. For example, mutation/deletion of *TP53* (a cell cycle controlling gene) has been observed in more than 50% of human cancers (Knudson, 2001), and high mutation rates and increased methylation levels have been detected for *RASSF1A*, *KRAS*, *BRAF*, and *MGMT* genes in CRC (Grady and Markowitz, 2002; Suehiro et al., 2008; Dworkin et al., 2009). Abnormal modifications of the *APC* gene have been associated with very early CRC stages (Suehiro et al., 2008) and aberrant alterations of *MLH1* and *MCC* genes found in hereditary and sporadic forms of CRC, respectively (Fukuyama et al., 2008). In Listing 3, we are interested in querying the EpiGeNet database for the highest interdependency between molecular events in carcinoma, based on conditional relationship data from StatEpigen. The objective is to explore potential hub events in CRC initiation. Results from this query (Table 2) indicate that *MLH1* hypermethylation, *JCVT* gene expression, and *CDKN2A:P16* hypermethylation are the three most frequently observed

Table 2. Epigenetic and Genetic Events (Top Three Results) Observed to be Interdependent with Other Molecular Events in Colon Carcinoma

Molecular event	*Neighbor no.*	*Neighbor set*
MLH1 HYPER-METH_CPG	52	Mutation: KRAS, MSH2, BRAF, APC, MLH1 Hypermethylation: MLH1, PLEKHC1, SOX7, C13ORF21, FLJ41549, PAPLN, ADAMTS19, FLJ37464, LRRC4, NPHS2, BMP3, MED12L, SLC30A10, EVL, DPYSL3, LYPD1, KCNK13, NELL2, SLC30A3, GDF7, NRG2, CLGN, CBS, KIT, FBXL7, ST3GAL1, TCF7L1, LOC283887, IMAGE5728979, CHFR, CDKN2A:P14, CDKN2A:P16, CRABP1, MGMT, APC, PTPRO, HLTF, SFRP1, RUNX3, TIMP3, DKK1 Gene expression: CDKN1A, MGMT, MLH1, JCVT; LOH: MLH1, TP53.
JCVT GENE_EXPRESSION	28	Mutation: KRAS, BRAF, PIK3CA Hypermethylation: MLH1, CHFR, CDKN2A:P14, CDKN2A:P16, CACNA1G, CRABP1, NEUROG1, MGMT, APC, HIC1, RUNX3, TIMP3, RARB, PTEN, WRN, IGF2, SOCS1, IGFBP3, APBA1, APBA3 Gene expression: TP53, CDKN1A, PTGS2, CTNNBIP1; LOH: RUNX1.
CDKN2A:P16 HYPER-METH_CPG	28	Mutation: KRAS, BRAF, PIK3CA Hypermethylation: RASSF1A, MLH1, CDKN2A:P14, CRABP1, MGMT, APC, MCC, HLTF, SLIT2, RUNX3, TIMP3, DAPK1 Gene expression: CDKN2A:P16, TP53, CDKN1A, CDKN1B, CCND1, PTGS2, MARCO, JCVT, FPGS Activation: CTNNB1 Polymorphism: MTHFR, MTR, MTRR.

LOH, loss of heterozygosity.

interdependent events (based on the genetic–epigenetic phenomena in carcinoma, curated from the literature, and stored in the StatEpigen database).

Listing 3. *Cypher query to identify molecular event neighborhood (in terms of connected nodes) in colon carcinoma:*

```
MATCH (m:MolecularEvent)-[:CARCINOMA]-(m2:MolecularEvent)
WITH m.GeneSymbol +" "+ m.EventName as MolecularEvent, collect(distinct
m2.GeneSymbol +" "+ m2.EventName) as NeighbourSet, count(distinct m2.GeneSymbol +" "+
m2.EventName) as NeighbourNo
RETURN MolecularEvent, NeighbourNo, NeighbourSet
ORDER BY NeighbourNo DESC
LIMIT 3
```

4.3. The most plausible paths (with highest incidence) connecting specific key molecular events (e.g., KRAS *mutation and* APC *hypermethylation) in e.g., adenoma*

We are interested in querying for occurrence of the most plausible paths (in terms of the connected graph of conditional relationships between molecular events implicated in CRC) that connect *KRAS* mutation and *APC* hypermethylation events [known to be important in CRC development (Grady and Markowitz, 2002; Suehiro et al., 2008)]. The Cypher query is given in Listing 4, and results are shown in Figure 3.

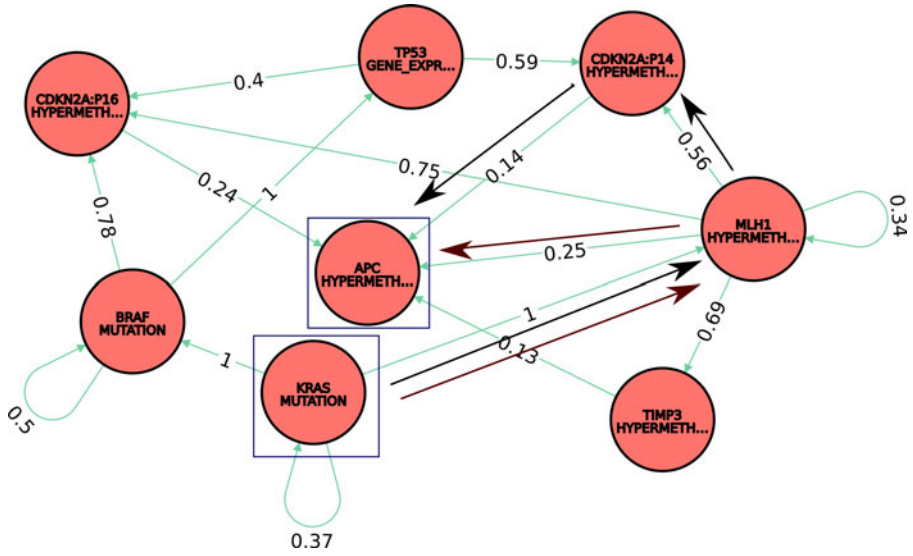

FIG. 3. Subnetwork with the most plausible paths between *KRAS* mutation and *APC* hypermethylation (blue contour), two molecular modifications known to be important in CRC. Legend: nodes: epigenetic and genetic events (e.g., *APC* hypermethylation, *KRAS* mutation); edges: colon carcinoma (i.e., the phenotype where the molecular signals have been observed). Information on conditional relationship probabilities of the molecular events is stored and displayed as edge attribute. Two examples of plausible pathways are highlighted in figure: Path1: *KRAS* mutation -> *MLH1* hypermethylation -> *APC* hypermethylation (brown arrows) and Path2: *KRAS* mutation -> *MLH1* hypermethylation -> *CDKN2A:p16* hypermethylation -> *APC* hypermethylation (black arrows).

Specifically, the query returns the top 10 most plausible pathways (based on a pathway overall score), composed of a maximum of five conditional molecular relationships connecting the two molecular signals of interest, *KRAS* mutation, and *APC* hypermethylation—highlighted by the blue contour in Fig. 3. For example, one such plausible pathway is

Path1: *KRAS* mutation -> *MLH1* hypermethylation -> *APC* hypermethylation (marked with brown arrows in Fig. 3). The overall score of Path1 indicates the probability of *APC* hypermethylation given *MLH1* hypermethylation and *KRAS* mutation, that is,

$$\text{Score}_{\text{Path1}} = \text{P}(\textit{APC}\ \text{HYPERMETH\_CPG} |\ \textit{MLH1}\ \text{HYPERMETH\_CPG and}\ \textit{KRAS}\ \text{MUTATION}) \quad (6)$$

Given that the conditional probabilities were measured independently, expression (Equation 6) can be written as a product of conditional probabilities as follows:

$$\begin{aligned}\text{Score}_{\text{Path1}} &= \text{P}(\textit{APC}\ \text{HYPERMETH\_CPG} |\ \textit{MLH1}\ \text{HYPERMETH\_CPG}) * \\ &\quad \text{P}(\textit{MLH1}\ \text{HYPERMETH\_CPG} |\ \textit{KRAS}\ \text{MUTATION}) = 0.25 * 1 = 0.25\ \text{(based on Fig. 3).}\end{aligned} \quad (7)$$

Similarly, an another pathway example (black arrows in Fig. 3) is

Path2: *KRAS* mutation -> *MLH1* hypermethylation -> *CDKN2A:p16* hypermethylation -> *APC* hypermethylation, with the overall score

$$\begin{aligned}\text{Score}_{\text{Path2}} &= \text{P}(\textit{APC}\ \text{HYPERMETH\_CPG} |\ \textit{CDKN2A:p16}\ \text{HYPERMETH\_CPG and}\ \textit{MLH1} \\ &\quad \text{HYPERMETH\_CPG and}\ \textit{KRAS}\ \text{MUTATION}) \\ &= \text{P}(\textit{APC}\ \text{HYPERMETH\_CPG} |\ \textit{CDKN2A:p16}\ \text{HYPERMETH\_CPG}) \\ &\quad * \textit{P}(\textit{CDKN2A:p16}\ \text{HYPERMETH\_CPG} |\ \textit{MLH1}\ \text{HYPERMETH\_CPG}) \\ &\quad * \textit{P}(\textit{MLH1}\ \text{HYPERMETH\_CPG} |\ \textit{KRAS}\ \text{MUTATION}) = 0.24 * 0.75 * 1 = 0.18 \\ &\quad \text{(based on Fig. 3).}\end{aligned} \quad (8)$$

Thus, the pathway overall score is computed as a product of the constituent conditional relationship probabilities. The ''reduce'' command in Listing 4 (below) computes automatically the scores of plausible pathways and returns the top 10 plausible pathways according to their score values. The maximum number of steps to be included in the plausible pathway (= 5 in the current case) and the ranked list (LIMIT = 10 in the current example) can be changed to other values if required.

Listing 4. *Cypher query to identify the first 10 most plausible pathways for a maximum of 5 conditional relationships between* KRAS *mutation and* APC *hypermethylation in carcinoma.*

```
MATCH p=(m1:MolecularEvent)-[:CARCINOMA*..5]-> (m2:MolecularEvent) where
m1.GeneSymbol="KRAS" and m1.EventName="MUTATION" and m2.GeneSymbol="APC"
and m2.EventName="HYPERMETH_CPG"
RETURN p AS plausiblePath, REDUCE(score=1, r in relationships(p) |
score*r.CondProbValue) AS totalScore
ORDER BY totalScore DESC
LIMIT 10
```

5. CONCLUSIONS AND FUTURE WORK

In comparison with relational databases, the graph database approach facilitates integration of *heterogeneous* and *highly connected* biodata and offers a natural representation of relationships among various concepts specific to biological systems (Have and Jensen, 2013). In addition, inspection of the diversity of the biological context, including traversal exploration in networks, identification of key elements (hubs)

within systems, and creation of modules to integrate concepts with high degree of similarity (e.g., based on common features, functions, and associations), can benefit from the use of graph-based algorithms. Consequently, results from this type of analysis can help generation of new hypotheses (linking diverse and differently structured concepts) that would be more difficult to formulate without the use of graph-based approaches. In this article, we have described EpiGeNet, a graph database that integrates data on genetic–epigenetic interdependencies observed at different pathological levels in CRC development. First, we used the Cypher language to query differences between polyps and adenoma phenotypes with respect to molecular modifications. Results indicate a set of 43 genetic–epigenetic conditional relationships, with 26 such relationships having probability of occurrence = $T > 0.60$. These events can be explored further to facilitate interpretation and identification of the mechanisms that differentiate between healthy phenotypes and those specific to CRC initiation. In Listings 3 and 4, details were queried on highly connected conditional molecular changes in carcinoma and on the most plausible paths connecting two major events.

Probabilistic computational models, reliant on available StatEpigen data (StatEpigen url, n.d.), were developed to investigate the most plausible pathways for cancer progression, based on genetic–epigenetic modifications at different stages of colorectal tumor. Variants of these models have also been used previously for exploration of DNA methylation dynamics during CRC initiation and progression [e.g., Roznovăţ and Ruskin (2013) and Barat and Ruskin (2015)]. The results from the EpiGeNet model adaptation can be incorporated into future computational and mathematical models of CRC development, with the added benefits of the combined probability basis for different event types (including those classified as self-aware or consecutive for same node) and graph algorithms for investigation of the gene connections and events involvement.

We have presented also a complementary graph database framework for integration of multiple heterogeneous biological types (Lysenko et al., 2016). This framework, which integrates data from major public resources [including DisGeNet (Bauer-Mehren et al., 2011), DrugBank (Knox et al., 2011), UniProt (Consortium, 2015)] and creates associations between different concepts (e.g., drugs-proteins-diseases), can be complemented by the EpiGeNet database for exploration in the context of human colonic disease.

While EpiGeNet has been developed initially using data available from the publicly accessible StatEpigen database (Barat and Ruskin, 2010), which indicates conditionality for epigenetic–genetic events in CRC, the future aim is integration of data on causality in molecular signals in CRC (i.e., the order in which these events occur). To achieve these further steps, text mining approaches will be utilized together with the Biological Expression Language (BEL) (BEL url, n.d.) to extend EpiGeNet with contextual information on CRC development (extracted from peer-reviewed publications) and enhance current available information.

ACKNOWLEDGMENTS

Authors would like to acknowledge access to the Neo4j 2.3.1 framework and the StatEpigen database. They would also like to thank the team of the CNRS/IN2P3 Computing Centre, Mr. J. Bussery, Mr. B. Guillon, Dr. G. Marchetti, and Dr. G. Rahal for their support in deploying and accessing the Neo4j graph database.

FUNDING

This work has been supported by the Innovative Medicines Initiative Joint Undertaking under grant agreement no. IMI 115446 (eTRIKS), resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies. A.L. and C.J.R. acknowledge support from the BBSRC through their strategic funding of Rothamsted Research.

AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

REFERENCES

Availability of the database: https://github.com/ibalaur/EpiGeNet.git. The queries described further can be explored at the following url: https://diseaseknowledgebase.etriks.org/epigenet/browser. Note: the ''AUTO-COMPLETE'' flag (Neo4j browser) was set to OFF when running the Cypher queries presented in this paper.

Alegría-Torres, J.A., Baccarelli, A., and Bollati, V. 2011. Epigenetics and lifestyle. *Epigenomics* 3, 267–277.

AllegroGraph url [www Document]. n.d. AllegroGraph RDFStore Web 30s database. Available at: http://franz.com/agraph/allegrograph Accessed January 7, 2016.

Allis, C.D., Jenuwein, T., Reinberg, D., and Caparros, M.-L. 2007. *Epigenetics*. CSHL Press.

Azad, N., Zahnow, C.A., Rudin, C.M., et al. 2013. The future of epigenetic therapy in solid tumours—Lessons from the past. *Nat. Rev. Clin. Oncol.* 10, 256–266.

Barat, A., and Ruskin, H.J. 2010. A manually curated novel knowledge management system for genetic and epigenetic molecular determinants of colon cancer. *Open Colorectal Cancer J.* 3, 36–46.

Barat, A., and Ruskin, H.J. 2015. Comparative correlation structure of colon cancer locus specific methylation: Characterisation of patient profiles and potential markers across 3 array-based datasets. *J. Cancer* 6, 795–811.

Bauer-Mehren, A., Bundschus, M., Rautschka, M., et al. 2011. Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. *PLoS One* 6, e20284.

Baylin, S.B., and Jones, P.A. 2011. A decade of exploring the cancer epigenome—Biological and translational implications. *Nat. Rev. Cancer* 11, 726–734.

BEL url [www Document]. n.d. Biological expression language BEL OpenBEL. Available at: www.openbel.org/bel-expression-language Accessed January 7, 2016.

Bjornsson, H.T., Daniele Fallin, M., and Feinberg, A.P. 2004. An integrated epigenetic and genetic approach to common human disease. *Trends Genet.* 20, 350–358.

Bottomly, D., McWeeney, S.K., and Wilmot, B. 2015. HitWalker2: Visual analytics for precision medicine and beyond. *Bioinformatics* 32, 1253–1255.

Bowel Cancer Incidence Statistics [www Document]. n.d. Cancer Res. UK. Available at: www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/incidence Accessed November 17, 2015.

Cancer of the Colon and Rectum—SEER Stat Fact Sheets [www Document]. n.d. Available at: http://seer.cancer.gov/statfacts/html/colorect.html Accessed November 17, 2015.

Carthew, R.W., and Sontheimer, E.J. 2009. Origins and mechanisms of miRNAs and siRNAs. *Cell* 136, 642–655.

Cedar, H., and Bergman, Y. 2009. Linking DNA methylation and histone modification: Patterns and paradigms. *Nat. Rev. Genet.* 10, 295–304.

Consortium, T.U. 2015. UniProt: A hub for protein information. *Nucleic Acids Res.* 43, D204–D212.

Dhir, M., Montgomery, E.A., Glöckner, S.C., et al. 2008. Epigenetic regulation of WNT signaling pathway genes in inflammatory bowel disease (IBD) associated neoplasia. *J. Gastrointest. Surg.* 12, 1745–1753.

Dworkin, A.M., Huang, T.H.-M., and Toland, A.E. 2009. Epigenetic alterations in the breast: Implications for breast cancer detection, prognosis and treatment. *Semin. Cancer Biol.* 19, 165–171.

FlockDB url [www Document]. n.d. Twitter open-sources home its soc. Graph Gigaom. Available at: http://gigaom.com/2010/04/12/twitter-open-sources-the-home-of-its-social-graph Accessed January 7, 2016.

Forbes, S.A., Beare, D., Gunasekaran, P., et al. 2015. COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* 43, D805–D811.

Fukuyama, R., Niculaita, R., Ng, K.P., et al. 2008. Mutated in colorectal cancer, a putative tumor suppressor for serrated colorectal cancer, selectively represses β-catenin-dependent transcription. *Oncogene* 27, 6044–6055.

Ghildiyal, M., and Zamore, P.D. 2009. Small silencing RNAs: An expanding universe. *Nat. Rev. Genet.* 10, 94–108.

Grady, W.M., and Markowitz, S.D. 2002. Genetic and epigenetic alterations in colon cancer. *Annu. Rev. Genomics Hum. Genet.* 3, 101–128.

Have, C.T., and Jensen, L.J. 2013. Are graph databases ready for bioinformatics? *Bioinformatics* 29, 3107–3108.

He, X., Chang, S., Zhang, J., et al. 2008. MethyCancer: The database of human DNA methylation and cancer. *Nucleic Acids Res.* 36, D836–D841.

Henkel, R., Wolkenhauer, O., and Waltemath, D. 2015. Combining computational models, semantic annotations and simulation experiments in a graph database. Database 2015, bau130.

HRGRN url [www Document]. n.d. Welcome HRGRN graph search-empowered integrative database of Arabidopsis signaling transduction, metabolism and gene regulation networks. Available at: http://plantgrn.noble.org/hrgrn Accessed January 8, 2016.

InfiniteGraph url [www Document]. n.d. Infinity distributed graph database—Object. Available at: www.objectivity.com/products/infinitegraph Accessed January 7, 2016.

Ito, T. 2007. Role of histone modification in chromatin dynamics. *J. Biochem. (Tokyo)* 141, 609–614.

Johnson, D., Connor, A.J., McKeever, S., et al. 2014. Semantically linking in silico cancer models. *Cancer Inform.* 13, 133–143.

Judson, H., Stewart, A., Leslie, A., et al. 2006. Relationship between point gene mutation, chromosomal abnormality, and tumour suppressor gene methylation status in colorectal adenomas. *J. Pathol.* 210, 344–350.

Knox, C., Law, V., Jewison, T., et al. 2011. DrugBank 3.0: A comprehensive resource for ''Omics'' research on drugs. *Nucleic Acids Res.* 39, D1035–D1041.

Knudson, A.G. 2001. Two genetic hits (more or less) to cancer. *Nat. Rev. Cancer* 1, 157–162.
Lysenko, A., Roznovat, I.A., Saqi, M., et al. 2016. Representing and querying disease networks using graph databases. BioData Mining 9(1), 1. DOI:10.1186/s13040-016-0102-8.
Mattick, J.S., Amaral, P.P., Dinger, M.E., et al. 2009. RNA regulation of epigenetic processes. *BioEssays* 31, 51–59.
Neo4j url [www Document]. n.d. Neo4j Worlds Lead. Graph Database. Available at: http://neo4j.com Accessed January 7, 2016.
OrientDB url [www Document]. n.d. OrientDB—OrientDB Multi-model NoSQL DatabaseOrientDB multi-model NoSQL database. Available at: http://orientdb.com/orientdb Accessed January 7, 2016.
Roznovăţ, I.A., and Ruskin, H.J. 2013. A computational model for genetic and epigenetic signals in colon cancer. *Interdiscip. Sci. Comput. Life Sci.* 5, 175–186.
Sparksee url [www Document]. n.d. Sparsity-Technol. Sparksee High-Perform. Graph Database. Available at: http://sparsity-technologies.com Accessed January 7, 2016.
StatEpigen url [www Document]. n.d. StatEpigen. Available at: http://statepigen.sci-sym.dcu.ie Accessed January 7, 2016.
Stein, R.A. 2014. Epigenetic therapies—A new direction in clinical medicine. *Int. J. Clin. Pract.* 68, 802–811.
Suehiro, Y., Wong, C.W., Chirieac, L.R., et al. 2008. Epigenetic-genetic interactions in the APC/WNT, RAS/RAF, and P53 pathways in Colorectal Carcinoma. *Clin. Cancer Res.* 14, 2560–2569.

Address correspondence to:
Dr. Irina Balaur
European Institute for Systems Biology and Medicine (EISBM)
CIRI UMR CNRS 5308
CNRS-ENS-UCBL-INSERM
Université Claude Bernard
3e étage plot 2
50 Avenue Tony Garnier
69366 Lyon cedex 07
France

E-mail: ibalaur@eisbm.org