KLIK: Kajian Ilmiah Informatika dan Komputer
ISSN 2723-3898 (Media Online)
Vol 3, No 6, Juni 2023, Hal 719-727
DOI 10.30865/klik.v3i6.932
https://djournals.com/klik

Integrasi data Protein-Protein Interactions dan Pathway untuk Menentukan Score pada pathway Menggunakan Analisis Graf

Lailan Sahrina Hasibuan[1*], Ahmad Fariqi[1], Lilik Prayitno[2], Melly Br Bangun[3]

[1] FMIPA, Departemen Ilmu Komputer, Institut Pertanian Bogor, Bogor, Indonesia
[2] Balai Besar Pengujian Mutu dan Sertifikasi Obat Hewan, Bogor, Indonesia
[3] Fakultas Ilmu Pendidikan, Pendidikan Luar Sekolah, Universitas Negeri Medan, Medan, Indonesia
Email: [1,*]lailan.sahrina@apps.ipb.ac.id, [2]fariqiahmad@gmail.com, [3]lilikprayitno58@gmail.com, [4]mellybgn@unimed.ac.id
Email Penulis Korespondensi: lailan.sahrina@apps.ipb.ac.id

Abstrak−Perkembangan teknologi biologi molekuler menghasilkan omics data dalam jumlah besar. Integrasi omics data bermanfaat untuk analisis proses biologi pada level molekuler, seperti ekspresi protein, mekanisme obat terhadap penyakit, dan mekanisme pewarisan sifat. Penelitian ini bertujuan mengintegrasikan data biologi molekuler protein melalui protein-protein interactions (PPIs), pathway, module dan orthology, untuk menghitung pathway score. Perhitungan score menggunakan perhitungan degree pada konsep graf. Protein, pathway, module dan ortholog berperan sebagai node, sementara interaksi diantaranya sebagai edge. Selanjtnya, sesuai dengan konsep graf bahwa node dengan degree yang tinggi menyatakan node yang memiliki peran penting dalam suatu graf. Berdasarkan konsep ini, pathway yang paling penting terkait suatu protein adalah pathway dengan degree tertinggi pada multipartite graf yang dibentuk oleh PPIs, module, ortholog dan pathway. Keluaran dari penelitian ini berupa package pada bahasa R untuk melakukan integrasi data biologi molekuler protein, pathway, module dan orthology, selanjutnya menampilkan pathway yang paling berperan terhadap protein berdasarkan urutan score tertinggi. Package ini diuji menggunakan masukan protein Insulin (INS) dan Xanthine dehydrogenase (XDH). Hasil perhitungan score pada pathway untuk INS menghasilkan pathway dengan score tertinggi yaitu MAPK signaling pathway (0.18) jalur 1, Pathways in cancer (0.137) jalur 2, Ubiquitin mediated proteolysis (0.28) jalur 3. Input protein XDH menghasilkan pathway Purine metabolism (0.67) jalur 1, Metabolic pathways (0.48) jalur 2 dan Purine metabolism (0.23) jalur 3. Hasil ini dapat dimanfaatkan untuk enrichment analysis mengenai hubungan antara protein dan pathway.

Kata Kunci: Module; Multipartite Graph; Orthologue; Pathway Score; Protein-Protein Interactions (PPIs)

Abstract−The development of molecular biology technology produces large amounts of omics data. Integration of omics data is useful for the analysis of biological processes at the molecular level, such as protein expression, drug mechanisms against diseases, and mechanisms of inheritance. This study aims to integrate protein molecular biology data through protein-protein interactions (PPIs), pathways, modules and orthology, to calculate pathway scores. The score calculation uses the degree calculation on the graph concept. Proteins, pathways, modules and orthologs act as nodes, while the interactions between them act as edges. Furthermore, according to the concept of a graph, nodes with a high degree represent nodes that have an important role in a graph. Based on this concept, the most important pathway related to a protein is the pathway with the highest degree in a multipartite graph formed by PPIs, modules, orthologs and pathways. The output of this study is a package in the R language to integrate data on molecular biology of proteins, pathways, modules and orthology, then displays the pathways that have the most role in protein based on the order of the highest score. This package was tested using protein Insulin (INS) and Xanthine dehydrogenase (XDH) inputs. The results of calculating the score on the pathway for INS produced the pathway with the highest score, namely MAPK signaling pathway (0.18) lane 1, Pathways in cancer (0.137) lane 2, Ubiquitin mediated proteolysis (0.28) lane 3. XDH protein input produces Purine metabolism pathway (0.67) lane 1, Metabolic pathways (0.48) lane 2 and Purine metabolism (0.23) lane 3. These results can be used for enrichment analysis regarding the relationship between proteins and pathways.

Keywords: Module; Multipartite Graph; Orthologue; Pathway Score; Protein-Protein Interactions (PPIs)

1. PENDAHULUAN

Bioinformatika adalah multidisipliner ilmu, yaitu gabungan antara ilmu biologi molekuler dan ilmu komputer [1] [2]. Perkembangan teknologi biologi molekuler telah menghasilkan data biologi molekuler dalam jumlah besar dan kompleks. Ilmu komputer diperlukan untuk menjawab permasalahan tentang data biologi molekuler menggunakan teknik algoritmik dan komputasi [3].

Diantara data biologi molekuler adalah *omics data*. *Omics data* terkait protein disebut *proteomics* yang tersedia pada repositori basis data *Search Tool for the Retrieval of Interacting Genes/Proteins* (STRING-db) [4], data gen disebut *genomics* yang tersedia di repositori basis data PubChem [5], dan data terkait *pathway* tersedia di repositori basis data *Kyoto Encyclopedia of Genes and Genomes* (KEGG) [6]. Data biologi molekuler atau data bioinformatika saling berhubungan satu sama lain. Data bioinformatika tersebut perlu diintegrasikan dan dianalisis untuk keperluan *enrichment analysis* guna mendapatkan informasi penting yang dibutuhkan manusia, seperti mengetahui gambaran ekspresi protein, mekanisme obat terhadap suatu penyakit, dan pewarisan sifat.

Protein adalah komponen utama organisme hidup dan kelas molekul yang sangat penting dalam setiap proses di dalam sel [7]. Sebagian besar fungsi protein adalah berinteraksi dengan molekul lain, terutama berinteraksi dengan sesama protein, juga dengan objek biologi molekuler lain, seperti *pathway*, *module*, dan *orthology* [8]. Interaksi suatu protein dengan protein lainnya dikenal dengan istilah *Protein-Protein Interactions* (PPIs). PPIs terjadi dalam banyak proses biologis dan berbagai penyakit. PPIs perlu diintegrasikan dan dianalisis dengan objek bioinformatika lain, seperti *pathway*, *module*, dan *orthology* untuk mengungkap dan menganalisis lebih lanjut tentang fungsi protein. Selain itu perthitungan *score* pada *pathway* juga diperlukan untuk mengungkap hubungan antara protein dengan *pathway*.

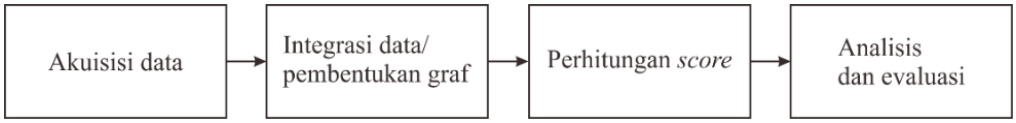
Penelitian sebelumnya yang membahas mengenai integrasi dan perhitungan *score* pada *omics data* telah banyak dilakukan, di antaranya adalah penelitian tentang integrasi dan perhitungan *score Single Nucleotide Polymorphisms* (SNPs) dengan gen dan *pathway* yang dilakukan oleh Lamparter *et al.* [9] dan prediksi PPIs dengan algoritma K-Nearest Neighbors yang dikombinasikan dengan Conjoint Triad (CT), Auto Covariance (AC) and Local Descriptor (LD) yang dilakukan oleh Yuanmiao Gui dan Xue Wang [7]. Selanjutnya, [10] melakukan scoring pada pathway terkait dua buah protein berdasarkan banyaknya pathway yang beririsan antara dua protein tersebut. Namun, integrasi data tentang PPIs, *pathway*, *module*, dan *orthology* belum dilakukan. Selain itu, penelitian tentang integrasi dan pergitungan *score* pada *omics data* masih perlu untuk terus dilakukan guna menghasilkan pengetahuan baru dan menguraikan sistem biologi molekuler yang kompleks [11].

Hal inilah yang melatarbelakangi penelitian ini. Maka pada penelitian ini dilakukan integrasi basis data bioinformatika yang berfokus pada integrasi data PPIs dengan *module*, *pathway*, dan *orthology*. Selain itu dilakukan perhitungan *score* pada *pathway* untuk melihat *pathway* paling berpengaruh. Perhitungan *score pathway* dilakukan melalui tiga jalur berdasarkan data *module* dan *orthology* dengan menghitung nilai *degree*. Penelitian ini menggunakan data masukan berupa Protein Insulin (INS) dan Xanthine dehydrogenase (XDH) untuk melihat *pathway* yang berhubungan dengan protein tersebut. Protein INS dipilih karena berkaitan dengan penyakit diabetes mellitus [12]. Sedangkan protein XDH dipilih karena berkaitan dengan penyakit gout atau lebih dikenal dengan penyakit asam urat [13]. Penyakit diabetes mellitus dan gout termasuk penyakit yang banyak diderita oleh masyarakat Indonesia.

Hasil penelitian ini berupa *package* untuk mengintegrasikan data biologi molekuler mengenai PPIs, *pathway*, *module*, dan *orthology*, serta melakukan perhitungan *score pathway* dengan menghitung nilai *degree.* Hasil penelitian ini dapat dimanfaatkan untuk *enrichment analysis* yang berguna untuk memahami hubungan antara protein dan *pathway* dan mendapatkan informasi lebih lanjut yang dibutuhkan oleh manusia.

2. METODOLOGI PENELITIAN

Tahapan metode penelitian terdiri atas akuisisi data, integrasi data/ pembentukan draf, perhitungan *score*, dan analisis dan evaluasi, seperti ditunjukkan pada Gambar 1.

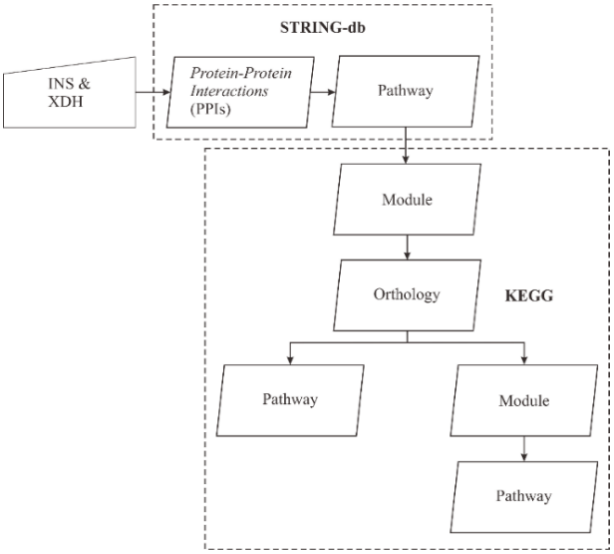


Gambar 1. Tahapan penelitian

2.1 Akuisisi Data

Akuisisi data adalah tahap pengambilan data dari repositori basis data STRING-db dan KEGG. Tahap ini dilakukan menggunakan dua metode, yaitu API untuk akuisisi data dari STRING-db dan *web scraping* untuk akuisisi data dari KEGG. Hasil akuisisi data berupa data PPIs, *pathway*, *module*, dan *orthology*, disimpan dalam *array* dalam format json.

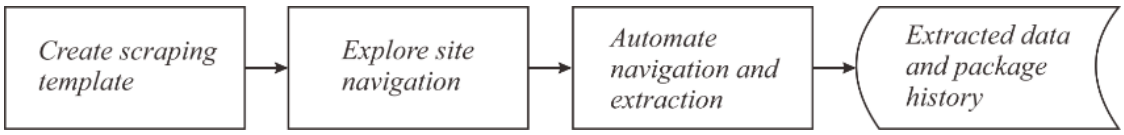
Tahap akuisisi data menggunakan dua metode, pertama akusisi data menggunakan *Application Programming Interface* (API), kedua teknik *web scraping.* Data yang menjadi masukan adalah data nama gene dari protein Insulin (INS) dan Xanthine dehydrogenase (XDH). Akuisisi data menggukanan API dilakukan untuk mengambil data PPIs dan *pathway* dari STRING-db. Sedangkan akuisi data menggunakan teknik *web scraping* dilakukan untuk mengambil data *module*, *pathway*, dan *orthology* dari KEGG. Tahapan akuisisi data dapat dilihat pada Gambar 2.



Gambar 2. Tahapan akuisisi data

a. *Application Programming Interface* (API)
API memungkinkan akuisisi data yang diinginkan tanpa menggunakan *user interface* dari sebuah halaman web. Alasan digunakannya API adalah karena data yang dihasilkan oleh STRING-db bersifat dinamis dan proses akuisisi data menggunakan API lebih cepat dan mudah karena disediakan secara legal oleh STRING-db.

b. *Web scraping*
Web scraping adalah proses pengambilan sebuah dokumen semi terstuktur dari internet, umumnya berupa halaman web dalam bahasa *markup* seperti HTML atau XHTML, dan menganalisis dokumen dari halaman tersebut untuk diambil data tertentu untuk digunakan bagi kepentingan lain. Teknik *web scraping* digunakan karena KEGG tidak menyediakan API. *Web scraping* memiliki empat langkah, yaitu: *Create Scraping Template*, mempelajari struktur dokumen HTML dari halaman web. Struktur dokumen HTML dipelajari untuk mengetahui *tag* HTML yang mengapit informasi yang diambil dan teknik pengambilan informasi tersebut. *Explore Site Navigation*, Peneliti mempelajari teknik navigasi pada halaman web. Teknik navigasi diperlukan untuk diaplikasikan pada program. *Automate Navigation and Extraction*, Berdasarkan informasi yang didapatkan pada langkah satu dan dua, program dibuat untuk mengotomatisasi pengambilan informasi dari halaman web yang ditentukan. *Extracted Data and Package History*, Informasi yang didapatkan dari langkah tiga disimpan dalam tabel-tabel penyimpanan data [14].
Langkah-langkah web scraping dapat dilihat pada Gambar 3.



Gambar 3. Alur *web scraping*

2.2 Integrasi data/ pembentukan graf

Tahapan ini adalah tahapan untuk mengintegrasikan data keluaran STRING-db dan KEGG, serta menyimpan data keluaran tersebut dalam bentuk graf. Integrasi data dilakukan dengan menghubungkan hasil akuisisi data bioinformatika dari STRING-db dengan KEGG.

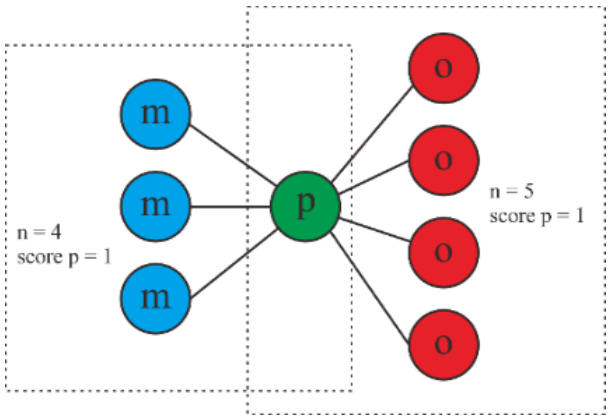
2.3 Perhitungan *score*

Perhitungan *score* dilakukan dengan menghitung *degree*. Perhitungan *score* dengan menggunakan *degree* dilakukan untuk menentukan *pathway* yang paling berpengaruh. Persamaan (1) merupakan persamaan yang digunakan untuk mendapatkan nilai *degree* [15].

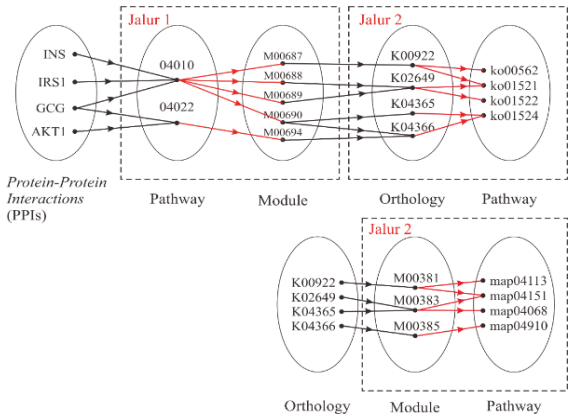
$$Cd(i) = \frac{d(i)}{n-1} \quad (1)$$

Keterangan:
$Cd(i)$ = *degree* suatu *node* ke-i.
$d(i)$ = banyaknya *edge* yang terbentuk antara *node* ke-i dengan *node* lain.
n = jumlah *node*.

Node merepresentasikan data *pathway*, *module*, atau *orthology*. *Edge* merepresentasikan hubungan antara *pathway* dan *module* atau *pathway* dan *orthology*. Ilustrasi perhitungan *score* dapat dilihat pada Gambar 4.



Gambar 4. Ilustrasi perhitungan score



Gambar 4. Jalur perhitungan *score pathway*

Perhitungan *score pathway* dilakukan melalui tiga jalur, yaitu jalur 1, jalur 2 dan jalur 3. Jalur 1 melakukan perhitungan *score pathway* berdasarkan *module*. Pada jalur pertama, data *pathway* diambil dari STRING-db dan data *module* diambil dari KEGG. Jalur 2 melakukan perhitungan *score pathway* berdasarkan *orthology*. Data *pathway* dan *orthology* diambil dari KEGG. Jalur 3 melakukan perhitungan *score pathway* berdasarkan *module*. Perbedaan antara jalur

satu dengan jalur tiga adalah data *pathway* pada jalur satu diambil dari STRING-db, sedangkan pada jalur 3 data *pathway* diambil dari KEGG. Ketiga jalur perhitungan *score pathway* dapat dilihat pada Gambar 5.

2.4 Analisis dan evaluasi

Analisis dan evaluasi dilakukan dengan mengkaji hasil perhitungan *score* pada *pathway* dari tiga jalur terhadap studi literatur berkaitan protein INS dan XDH. Konsep *degree* pada *graph mining* menyebutkan bahwa nilai *degree* yang semakin tinggi pada suatu node menandakan node tersebut berperan penting pada graf [16].

3. HASIL DAN PEMBAHASAN

3.1 Akuisisi Data

Akuisisi data adalah tahap pengambilan data dari repositori basis data STRING-db dan KEGG. Tahap ini dilakukan menggunakan dua metode, yaitu API untuk akuisisi data dari STRING-db dan *web scraping* untuk akuisisi data dari KEGG. Hasil akuisisi data berupa data PPIs, *pathway*, *module*, dan *orthology*, disimpan dalam *array* dalam format json. Penelitian ini menggunakan data masukan berupa protein Insulin (INS) dan Xanthine dehydrogenase (XDH).

3.1.1 *Application Programming Interface* (API)

API digunakan untuk melakukan akuisisi data PPIs dan *pathway* dari STRING-db. Data yang pertama kali diakuisisi dari STRING-db adalah PPIs. Berikut adalah API untuk akuisisi data PPIs, https://string-db.org/api/json/network?identifiers=[*identifiers*]. Identifier yang digunakan sebagai masukan untuk mendapkan data PPIs adalah nama gen dari protein. Misalnya pada penelitian ini untuk mendapatkan PPIs dari protein Insulin dan Xanthine dehydrogenase *identifier* yang digunakan adalah nama gen dari protein Insulin yaitu INS dan nama gen dari Xanthine dehydrogenase yaitu XDH. Nama gen dari protein tersebut diperoleh dari respositori basis data *The Universal Protein Resource* (UniProt) [17].

Akuisisi data PPIs menggunakan API menghasilkan 13 atribut, yaitu atribut stringId_A, stringId_B, preferredName_A, preferredName_B, ncbiTaxonId, score, nscore, fscore, pscore, ascore, dscore, dan tscore. Dari 13 atribut yang dihasilkan hanya dua atribut yang merupakan data PPIs, yaitu atribut preferredName_A dan preferredName_B. Selain atribut preferredName_A dan preferredName_B bukan data PPIs sehingga tidak diperlukan. Contoh data PPIs yang diperoleh dari protein Insulin dan Xanthine dehydrogenase dapat dilihat pada Gambar 6 dan Gambar 7.

a

stringId_A	stringId_B	b preferredName_A	preferredName_B	ncbiTaxonId	score	nscore	fscore	pscore	ascore	escore	dscore	tscore
ENSP00000250971	ENSP00000225577	INS	RPS6KB1	9606	0.987	0	0	0	0	0.9	0.882	
ENSP00000263967	ENSP00000225577	PIK3CA	RPS6KB1	9606	0.997	0	0	0	0.061	0.527	0.9	0.944
ENSP00000263967	ENSP00000250971	PIK3CA	INS	9606	0.989	0	0	0	0	0	0.9	0.895
ENSP00000265171	ENSP00000225577	EGF	RPS6KB1	9606	0.754	0	0	0	0.049	0	0	0.752
ENSP00000265171	ENSP00000250971	EGF	INS	9606	0.99	0	0	0	0	0	0.8	0.954
ENSP00000265171	ENSP00000263967	EGF	PIK3CA	9606	0.988	0	0	0	0	0	0.9	0.892
ENSP00000270202	ENSP00000225577	AKT1	RPS6KB1	9606	0.97	0	0	0	0.05616	0.106	0.633	0.9	0.12519
ENSP00000270202	ENSP00000250971	AKT1	INS	9606	0.996	0	0	0	0	0	0	0.9	0.962
ENSP00000270202	ENSP00000263967	AKT1	PIK3CA	9606	0.999	0	0	0	0	0.094	0.551	0.9	0.99
ENSP00000270202	ENSP00000265171	AKT1	EGF	9606	0.943	0	0	0	0	0.049	0	0	0.942

Gambar 5. Data hasil API dengan masukan INS. (a) Nama atribut yang menyimpan data PPIs. (b) Data PPIs

a

stringId_A	stringId_B	b preferredName_A	preferredName_B	ncbiTaxonId	score	nscore	fscore	pscore	ascore	escore	dscore	tscore
ENSP00000298556	ENSP00000238018	HPRT1	GDA	9606	0.928	0.059	0	0	0	0	0.9	0.303
ENSP00000312304	ENSP00000286479	TPMT	NAT2	9606	0.442	0	0	0	0	0	0	0.442
ENSP00000312304	ENSP00000298556	TPMT	HPRT1	9606	0.958	0	0	0	0	0	0.9	0.598
ENSP00000342007	ENSP00000286479	CYP1A2	NAT2	9606	0.971	0	0	0	0.09	0	0.9	0.714
ENSP00000354532	ENSP00000238018	PNP	GDA	9606	0.987	0.309	0	0	0.058	0	0.941	0.723
ENSP00000354532	ENSP00000298556	PNP	HPRT1	9606	0.983	0.309	0	0	0.054	0	0.941	0.62
ENSP00000363832	ENSP00000238018	AOX1	GDA	9606	0.853	0.505	0	0.343	0.131	0	0.536	0.053
ENSP00000363832	ENSP00000261326	AOX1	MOCOS	9606	0.954	0	0	0	0.11	0	0	0.95
ENSP00000363832	ENSP00000342007	AOX1	CYP1A2	9606	0.931	0	0	0	0.052	0	0.9	0.339
ENSP00000363832	ENSP00000354532	AOX1	PNP	9606	0.552	0	0	0	0.056	0	0.536	0.061

Gambar 6. Data hasil API dengan masukan XDH. (a) Nama atribut yang menyimpan data PPIs. (b) Data PPIs

Data PPIs yang diperoleh selanjutnya digunakan sebagai *identifier* untuk akuisisi data *pathway*. API untuk akuisisi data *pathway* adalah sebagai berikut, https://string-db.org/api/tsv/enrichment?identifiers=[*identifiers*]. Hasil akuisisi data *pathway* menghasilkan 10 atribut, yaitu atribut category, term, number_of_genes, ncbiTaxonId, inputGenes, preferredNames, p_value, fdr, bonferroni, dan description. Data *pathway* disimpan pada atribut term dengan nama *category* KEGG. Contoh data *pathway* dari hasil akuisisi data dapat dilihat pada Gambar 8 dan Gambar 9.

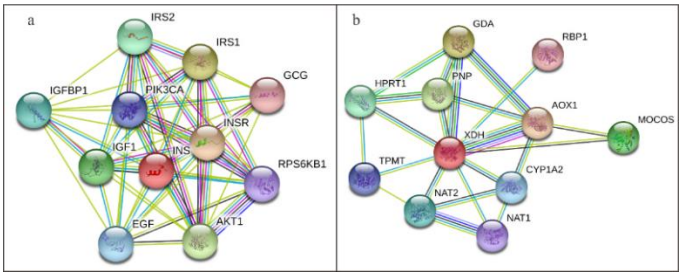
a category	b term	number_of_genes	ncbiTaxonId	inputGenes	preferredNames	p_value	
Process	GO:0044281	2	9606	NAT2,XDH	NAT2,XDH	0.00906	1	1
KEGG	01100	2	9606	NAT2,XDH	NAT2,XDH	0.00322	0.308	0.924	Metaboli
KEGG	00983	2	9606	NAT2,XDH	NAT2,XDH	4.32e-06		0.000619	
KEGG	00232	2	9606	NAT2,XDH	NAT2,XDH	4.78e-08		1.37e-05	
Component	GO:0005829	2	9606	NAT2,XDH	NAT2,XDH		0.0215	1	
Function	GO:0003824	2	9606	NAT2,XDH	NAT2,XDH		0.056	1	

Gambar 7. Data hasil API dengan masukan PPIs dari XDH. (a) Nama *category* untuk data pathway. (b) Data pathway

	category	term	number_of_genes	ncbiTaxonId	inputGenes	preferredNames	p_value
	Process	GO:0030335	2	9606	RPS6KB1,INS	RPS6KB1,INS	0.00025 0.23	1
	Process	GO:0008286	2	9606	RPS6KB1,INS	RPS6KB1,INS	5.27e-05	0.121
	Process	GO:0032270	2	9606	RPS6KB1,INS	RPS6KB1,INS	0.00338 1	1
a	Process	GO:0045927	2	9606	RPS6KB1,INS	RPS6KB1,INS	0.000117	0.18
	KEGG	04150	2	9606	RPS6KB1,INS	RPS6KB1,INS	8.18e-06	0.00235 0.00235
	KEGG	04066	2	9606	RPS6KB1,INS	RPS6KB1,INS	2.51e-05	0.00296 0.00721
	KEGG	04910	2	9606	RPS6KB1,INS	RPS6KB1,INS	4.13e-05	0.00296 0.0119
	KEGG	04152	2	9606	RPS6KB1,INS	RPS6KB1,INS	3.36e-05	0.00296 0.00963
	KEGG	04151	2	9606	RPS6KB1,INS	RPS6KB1,INS	0.000269	0.0154 0.0772
	Component	GO:0070013	2	9606	RPS6KB1,INS	RPS6KB1,INS	0.0341	1

Gambar 8. Data hasil API dengan masukan PPIs dari INS. (a) Nama *category* untuk data pathway. (b) Data pathway

Hasil akuisisi data PPIs dengan masukan INS menghasilkan sepuluh protein yang berinteraksi dengan INS, dan hasil akuisisi data PPIs dengan masukan XDH juga menghasilkan sepuluh protein yang berinteraksi dengan XDH. Visualisasi PPIs dengan masukan INS dan XDH dapat dilihat pada Gambar 10.



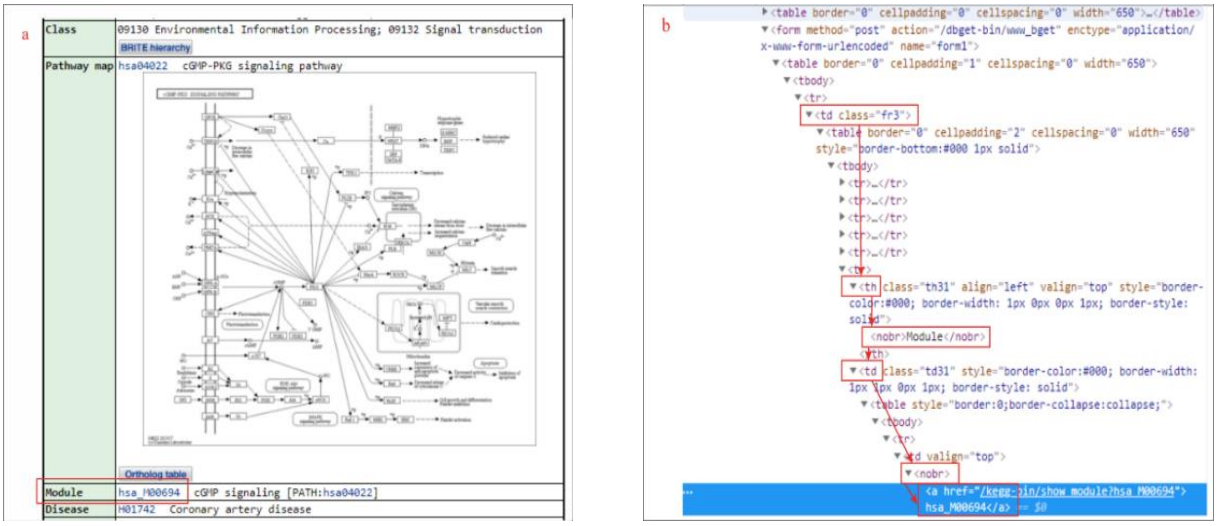
Gambar 9. Visualisasi PPIs dengan masukan adalah protein yang berwarna merah. (a) Visualisasi PPIs dengan masukan INS. (b) Visualisasi PPIs dengan masukan XDH

3.1.2 *Web scraping*

Teknik *web scraping* digunakan untuk akuisisi data *module*, *orthology*, dan *pathway* dari repositori basis data KEGG. Struktur halaman web KEGG disimpan di dalam tabel dan *tag* tertentu. Hasil empat tahap *web scrapping* basisdata KEGG adalah sebagai berikut:

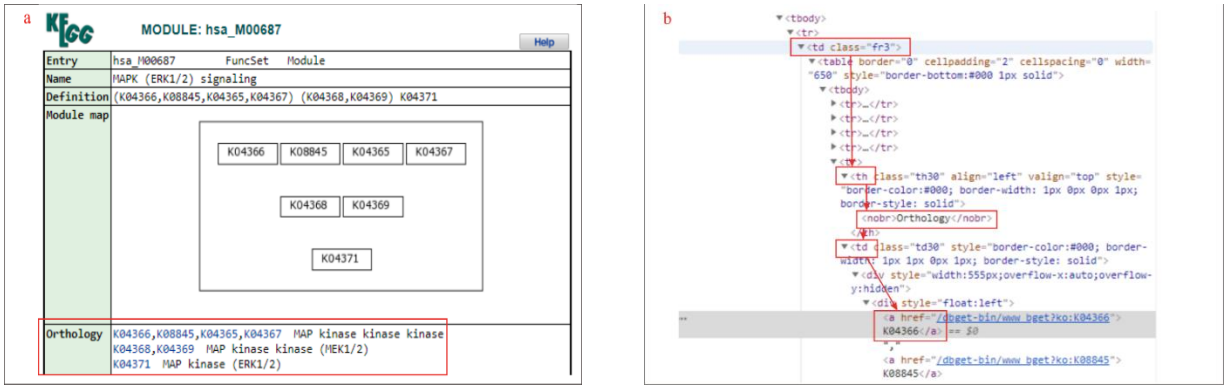
a. *Create Scrapping Template*

Secara umum *create scraping template* untuk akuisisi data pada halaman web KEGG adalah kegiatan menemukan *class* dari tabel penyimpanan data, mengambil semua *tag* <th> melalui *class* dari tabel penyimpanan data, menemukan *tag* <nobr> yang berisi nama data yang akan diambil, mengambil data melalui *tag* <nobr>, data *module* disimpan di dalam *tag* <a> yang bersarang pada *tag* <nobr>. Tampilan halaman dan struktur HTML halaman web KEGG *pathway* yang menyimpan data *module* dapat dilihat pada Gambar 11.



Gambar 10. KEGG pathway untuk akuisisi data module. (a) Tampilan halaman KEGG pathway. (b) Struktur halaman KEGG pathway dan *scraping template* untuk proses akuisisi data module pada KEGG pathway.

Data *orthology* disimpan di dalam *tag* <a>. Tampilan halaman dan struktur HTML halaman web KEGG *pathway* yang menyimpan data *orthology* dapat dilihat pada Gambar 12. Data *module* disimpan di dalam *tag* <a> yang bersarang pada *tag* <nobr>. Tampilan halaman dan struktur HTML halaman web KEGG *orthology* yang menyimpan data *module* dapat dilihat pada Gambar 13.



Gambar 11. KEGG module untuk akuisisi data *orthology*.(a) Tampilan halaman KEGG module. (b) Struktur halaman KEGG module dan *scraping template* untuk proses akuisisi data *orthology* pada KEGG modul.

Gambar 12. KEGG *orthology* untuk akuisisi data module.(a) Tampilan halaman KEGG *orthology*. (b) Struktur halaman KEGG *orthology* dan *scraping template* untuk proses akuisisi data module pada KEGG *orthology*.

b. *Explore Site Navigation*

Tahap e*xplore site navigation* dilakukan dengan mengidentifikasi dan memahami logika struktur URL. Berikut adalah struktur URL halaman web KEGG untuk akuisisi data: *Module* berdasarkan *pathway*: https://www.genome.jp/dbget-bin/www_get?pathway:hsa[*identifier*]. *Orthology* berdasarkan *module* dan *pathway* berdasarkan *module*: https:// https://www.genome.jp/dbget-bin/www_get?[*identifier*]. *Pathway* dan *module* berdasarkan orthology: http://www.genome.jp/dbget-bin/www_bget?ko:[*identifier*].

Setiap URL dari halaman web KEGG memiliki *identifier*. *Identifier* adalah data masukan yang nilainya berubah jika sebuah halaman web diobservasi.

c. *Automate Navigation and Extraction*

Langkah *automate navigation and extraction* dibuat sesuai dengan alur akuisisi data yang telah dibuat berdasarkan langkah *create scraping template.* Langkah *automate navigation and extraction* diawali dengan membuat *request* halaman web KEGG. Kemudian untuk mendapatkan data *module*, *orthology*, dan *pathway* digunakan html.parser dan beberapa *method* dari *package* BeautifulSoup. Data *module*, *orthology*, dan *pathway* didapatkan dengan mengekstrak tabel pada halaman web KEGG.

d. *Extracted Data and Package History*

Langkah terakhir dari *web scraping* adalah *extracted data and package history*. Data *module*, *orthology*, dan *pathway* yang telah berhasil diakuisisi disimpan di dalam tabel-tabel *array* json. Contoh potongan program untuk menyimpan data dapat dilihat pada Gambar 14, dan contoh hasil akuisisi data menggunakan teknik *web scraping* dapat dilihat pada Gambar 15.

```
# save data pathway & module
data = {
    "pathway": pathway_id,
    "module": i
    }
pathway_module.append(data)
```

Gambar 14. Potongan program untuk menyimpan data

```
orthology_pathway = [{'orthology': u'K04427', 'pathway': u'ko04310'},
    {'orthology': u'K04427', 'pathway': u'ko04380'},
    {'orthology': u'K04427', 'pathway': u'ko04520'},
    {'orthology': u'K04427', 'pathway': u'ko04620'},
    {'orthology': u'K04427', 'pathway': u'ko04621'},
    {'orthology': u'K04427', 'pathway': u'ko04622'}]
```

Gambar 135. Data hasil akuisisi data menggunakan *web scraping*

3.2. Integrasi Data/ Pembentukan Graf

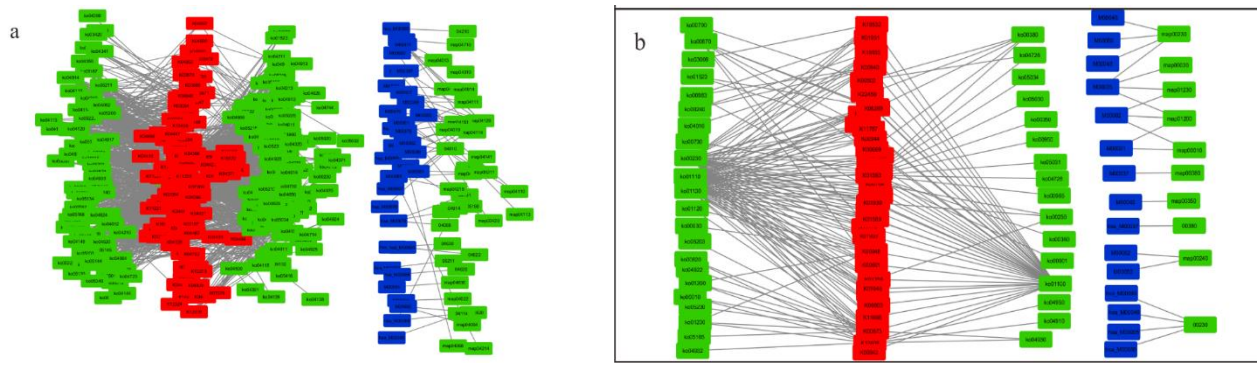
Tahap integrasi data dilakukan dengan menghubungkan data yang didapatkan dari hasil akuisisi data. Alur integrasi data dilakukan sesuai dengan alur akuisisi data yang telah dibuat. Alur integrasi data adalah sebagai berikut: integrasi PPIs dengan *pathway* pada STRING-db, integrasi *pathway* dengan *module* pada KEGG, integrasi *module* dengan *orthology* pada KEGG, integrasi *orthology* dengan *pathway* pada KEGG, integrasi *orthology* dengan *module* pada KEGG, dan integrasi *module* dengan *pathway* pada KEGG.

Data hasil integrasi disimpan dalam bentuk graf dalam format json. Graf yang dihasilkan disimpan di dalam *array* dalam format json, seperti ditunjukkan pada Gambar 16.

```
orthology_pathway = [{'orthology': u'K04427', 'pathway': u'ko04310'},
    {'orthology': u'K04427', 'pathway': u'ko04380'},
    {'orthology': u'K04427', 'pathway': u'ko04520'},
    {'orthology': u'K04427', 'pathway': u'ko04620'},
    {'orthology': u'K04427', 'pathway': u'ko04621'},
    {'orthology': u'K04427', 'pathway': u'ko04622'}]
```

Gambar 146. Graf dalam bentuk *array* json

Graf dalam bentuk *array* json kemudian dikonversi ke dalam format csv dan divisualisasikan menggunakan perangkat lunak Cytoscape versi 3.6.1. Hasil visualisasi graf dengan masukan INS dan XDH dapat dilihat pada Gambar 17.



Gambar 157. Visualisasi data hasil integrasi pathway, module, dan *orthology* dalam bentuk graf. Warna hijau adalah pathway, warna merah adalah *orthology*, dan warna biru adalah module. (a) Graf integrasi untuk protein INS. (b) Graf integrasi untuk protein XDH

3.3 Perhitungan *Score*

Perhitungan *score* dilakukan dengan menghitung nilai *degree. Pathway* dominan hasil perhitungan *score pathway* dengan masukan protein INS dari tiga jalur perhitungan dapat dilihat pada Tabel 1, dan *pathway* dominan hasil perhitungan *score pathway* dengan masukan protein XDH dari tiga jalur dapat dilihat pada Tabel 2.

Tabel 1. *Pathway* dominan hasil perhitungan *score pathway* dengan masukan protein INS

No	Jalur	Kode *pathway*	*Score*	Nama *pathway*
1	1	04010	0.18	MAPK *signaling pathway*
2	2	ko05200	0.137	*Pathways in cancer*
3	3	map04120	0.28	*Ubiquitin mediated proteolysis*

Tabel 2. *Pathway* dominan hasil perhitungan *score pathway* dengan masukan protein XDH

No	Jalur	Kode *pathway*	*Score*	Nama *pathway*
1	1	00230	0.67	*Purine metabolism*
2	2	ko01100	0.48	*Metabolic pathways*
3	3	map00230	0.23	*Purine metabolism*

3.4 Analisis dan Evaluasi

Konsep *degree* pada *graph mining* menyebutkan bahwa nilai *degree* yang tinggi dianggap sebagai node yang berperan penting pada graf tersebut. Dengan menggunakan konsep ini, dapat disimpulkan bahwa *pathway* yang paling penting untuk protein INS adalah MAPK *signaling pathway*, *Pathways in cancer*, dan *Ubiquitin mediated proteolysis*, serta *pathway* yang paling penting untuk protein XDH adalah *Purine metabolism* dan *Metabolic pathways* berdasarkan jalur 1, jalur 2, dan jalur 3.

Menurut J. Barletto Sousa Barros, *et. al.* dalam [18], insulin memiliki peranan besar dalam mengaktifkan MAPK signalling pathway, yang mana MAPK signalling pathway memiliki peranan penting dalam perkembangan penyakit diabetes. Hal ini dikonfirmasi kembali oleh Kurauti, *et. al.* dalam penelitiannya mengenai insulin dan *aging* bahwa insulin menstimulasi mitogenesis melalui peangaktifan mitogen-activated protein kinase (MAPK) pathway [19]. Selanjutnya,

penelitian (Poloz dan Stambolic) mengenai obesitas dan kanker mengungkapkan bahwa penyakit diabetes memiliki hubungan yang kuat terhadap obesitas. Penderita obesitas memiliki jaringan adiposa yang tidak normal dan ini memicu diabetes maupun kanker melalui respon tubuh terhadap insulin, baik untuk *insuline resistance* (IR) maupun *hyperinsulinemia* [20]. Sementara itu, penelitian [21] mengungkap hubungan antara insulin dan *Ubiquitin mediated proteolysis*. Sistem ubiquitin/protea-some (UPS) memengaruhi fungsionalitas IIS (insulin/insulin-like growth factor-1 (IGF-1) signaling) melalui jalur ubiquitylation yang dapat diinduksi yang mengatur internalisasi reseptor insulin/IGF-1, stabilitas target pensinyalan insulin/IGF-1 hilir, dan aktivitas reseptor inti sel untuk pengontrolan ekspresi gen.

Selanjutnya untuk protein XDH (Xanthine dehydrogenase), [22] menyebutkan XDH bahwa adalah enzim pembatas laju dalam katabolisme purin dengan mengubah hipoksantin menjadi xantin dan xantin menjadi asam urat. Ini menunjukkan hubungan yang erat antara XDH dan penyakit gout yang disebabkan tingginya kadar asam urat di dalam darah. Sementara hubungan antara XDH dan *Metabolic pathways* sangat luas, karena *Metabolic pathways* merupakan pathway yang mengatur metabolisme di dalam tubuh.

Hasil penelitian ini dapat dimanfaatkan untuk *enrichment analysis* yang berguna untuk memahami hubungan antara protein dan *pathway* dan mendapatkan informasi lebih lanjut yang dibutuhkan oleh manusia. *Enrichment analysis* yaitu analisis yang dilakukan untuk melihat proses atau temuan lain pada suatu proses biologi molekuler [23]. Pada penelitian ini digunakan protein INS yang berkaitan dengan penyakit diabetes mellitus dan protein XDH yang berkaitan dengan penyakit gout atau asam urat, maka *enrichment analysis* dilakukan untuk melihat temuan lain hubungan protein INS dengan penyakit atau proses lain dan hubungan protein XDH dengan penyakit atau proses lain.

4. KESIMPULAN

Penelitian ini menghasilkan *package* yang dapat digunakan untuk integrasi data biologi molekuler mengenai PPIs, *pathway*, *module*, dan *orthology*, serta perhitungan *score pathway* dengan menghitung nilai *degree*. Selain itu penelitian ini juga berhasil menganalisis hubungan antara protein dengan *pathway* yang selanjutnya dapat digunakan untuk melakukan *enrichment analysis* terkait protein dan *pathway*. Hasil perhitungan *score pathway* dengan masukan protein INS dan protein XDH dengan tiga jalur perhitungan menghasilkan lima *pathway* dominan. Masukan berupa protein INS menghasilkan *pathway* dominan MAPK signaling *pathway*, *Pathways in cancer*, dan *Ubiquitin mediated*. Selanjutnya, masukan berupa protein XDH menghasilkan *pathway Purine metabolism* dan *Metabolic pathways*. Bukti ilmiah mengenai hubungan antara protein masukan dan *pathway* yang dihasilkan oleh *package* menunjukkan bahwa konsep degree pada teori graph dapat diterapkan pada multi-grap data biologi molekuler.

REFERENCES

[1] A. Srivastava and A. Naik, "Big Data Analysis in Bioinformatics," *Adv. Bioinformatics*, vol. 2021, pp. 405–429, 2021.
[2] J. Gauthier, A. T. Vincent, S. J. Charette, and N. Derome, "A brief history of bioinformatics," *Brief. Bioinform.*, vol. 20, no. 6, pp. 1981–1996, Nov. 2019.
[3] M. C. Schatz *et al.*, "Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space," *Cell Genomics*, vol. 2, no. 1, pp. 1–13, 2022.
[4] D. Szklarczyk *et al.*, "The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D362–D368, 2017.
[5] S. Kim *et al.*, "PubChem 2019 update: Improved access to chemical data," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1102–D1109, 2019.
[6] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, "KEGG: New perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D353–D361, 2017.
[7] Y. Gui and X. Wang, "Application of K-nearest neighbors in protein-protein interaction prediction," *Highlights Sci. Eng. Technol.*, vol. 2, pp. 125–131, 2022.
[8] A. Fadli, W. A. Kusuma, Annisa, I. Batubara, and R. Heryanto, "Screening of potential Indonesia herbal compounds based on multi-label classification for 2019 coronavirus disease," *Big Data Cogn. Comput.*, vol. 5, no. 4, 2021.
[9] D. Lamparter, D. Marbach, R. Rueedi, Z. Kutalik, and S. Bergmann, "Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics," *PLoS Comput. Biol.*, vol. 12, no. 1, pp. 1–20, 2016.
[10] S. Kitsiranuwat, A. Suratanee, and K. Plaimas, "Multi-data aspects of protein similarity with a learning technique to identify drug-disease associations," *Appl. Sci.*, vol. 11, no. 7, p. 2914, Apr. 2021.
[11] W. Zhang, F. Li, L. Nie, and W. Zhang WeiwenZhang, "Integrating multiple 'omics' analysis for microbial biology: application and methodologies."
[12] P. S. Nugroho, N. A. Tianingrum, S. Sunarti, A. Rachman, D. S. Fahrurodzi, and R. Amiruddin, "Predictor risk of diabetes mellitus in Indonesia, based on national health survey," *Malaysian J. Med. Heal. Sci.*, vol. 16, no. 1, pp. 126–130, 2020.
[13] R. Afnuhazi, "Faktor - Faktor Yang Berhubungan Dengan Kejadian Asam Urat Pada Lansia (45 – 70 Tahun)," *Hum. Care J.*, vol. 4, no. 1, p. 34, 2019.
[14] G. Ginde *et al.*, "ScientoBASE: a framework and model for computing scholastic indicators of non-local influence of journals via native data acquisition algorithms," *Scientometrics*, vol. 108, no. 3, pp. 1479–1529, 2016.
[15] A. Jain and B. V. R. Reddy, "Optimal degree centrality based algorithm for cluster head selection in wireless sensor networks," *2014 Recent Adv. Eng. Comput. Sci. RAECS 2014*, pp. 6–8, 2014.
[16] Z. Huang *et al.*, "Network Pharmacology Approach to Uncover the Mechanism Governing the Effect of Simiao Powder on Knee Osteoarthritis," *Biomed Res. Int.*, vol. 2020, pp. 1–12, 2020.
[17] A. Bateman, "UniProt: A worldwide hub of protein knowledge," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D506–D515, 2019.

[18] J. Barletto Sousa Barros, R. da Silva Santos, and A. Adamski da Silva Reis, "Implication of the MAPK Signalling Pathway in the Pathogenesis of Diabetic Nephropathy," *EMJ Diabetes*, no. November, pp. 107–114, 2019.
[19] M. A. Kurauti, G. M. Soares, C. Marmentini, G. A. Bronczek, R. C. S. Branco, and A. C. Boschero, "Insulin and aging," *Vitam. Horm.*, vol. 115, pp. 185–219, Jan. 2021.
[20] Y. Poloz and V. Stambolic, "Obesity and cancer, a case for insulin signaling," *Cell Death Dis.*, vol. 6, no. 12, pp. e2037-11, 2015.
[21] V. Balaji, W. Pokrzywa, and T. Hoppe, "Ubiquitylation Pathways In Insulin Signaling and Organismal Homeostasis," *BioEssays*, vol. 40, no. 5, pp. 1–10, 2018.
[22] M. M. Chen *et al.*, "Xanthine dehydrogenase rewires metabolism and the survival of nutrient deprived lung adenocarcinoma cells by facilitating UPR and autophagic degradation," *Int. J. Biol. Sci.*, vol. 19, no. 3, pp. 772–788, 2023.
[23] S. Mubeen, C. T. Hoyt, A. Gemünd, M. Hofmann-Apitius, H. Fröhlich, and D. Domingo-Fernández, "The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling," *Front. Genet.*, vol. 10, no. November, pp. 1–13, 2019.