Pemfilteran kolaboratif mendalam untuk prediksi gen penyakit

Xiangxiang Zeng, *Anggota Senior, IEEE,* Yinglai Lin, Yuying He , Linyuan Lu, Xiaoping Min ¨*, dan Alfonso Rodr´ıguez-Paton´Pengutamaan yang akurat

Abstrak —dari gen penyakit potensial merupakan tantangan mendasar dalam penelitian biomedis. Berbagai algoritma telah dikembangkan untuk menyelesaikan masalah tersebut. Inductive Matrix Completion (IMC) adalah salah satu model paling andal untuk
kerangka kerjanya yang mapan dan kinerjanya yang superior dalam memprediksi asosiasi penyakit gen. Namun, metode IMC tidak mengekstrak fitur dalam secara hierarki, yang mungkin membatasi kualitas pemulihan. Dalam hal ini, arsitektur pembelajaran mendalam, yang memperoleh representasi tingkat tinggi dan menangani kebisingan dan pencilan yang disajikan dalam kumpulan data biologis berskala besar, dimasukkan ke dalam informasi samping gen dalam model Deep Collaborative Filtering (DCF) kami. Selanjutnya, karena kurangnya contoh negatif, kami juga mengeksploitasi formulasi pembelajaran Positive-Unlabeled (PU) untuk penyelesaian matriks peringkat rendah. Pendekatan kami mencapaijauh lebih
kinerja yangbaik dibandingkan metode canggih lainnya pada penyakit dari database Online Mendelian Inheritance in Man (OMIM). Pendekatan kami 10% lebih efisien daripada IMC standar dalam mendeteksi hubungan yang sebenarnya, dan secara signifikan mengungguli alternatif lain dalam hal metrik presisi-recall pada prediksi top-k. Selain itu, kami juga memvalidasi penyakit tanpa asosiasi gen yang diketahui sebelumnya dan asosiasi OMIM yang baru dilaporkan. Hasil eksperimen menunjukkan bahwa DCF masih memuaskan untuk pemeringkatan fenotipe penyakit baru serta hubungan pertambangan yang belum dijelajahi. Kode sumber dan datanya tersedia di
https://github.com/xzenglab/Deep-Collaborative-Filtering.

Istilah Indeks —Prediksi gen penyakit, Integrasi data, Pembelajaran mendalam, pembelajaran PU, Penyelesaian matriks.

✦

1ENDAHULUANP

adalah mengidentifikasi linkgen-penyakit yang nilai besar dalam

diagnosis dan pengobatan penyakit manusia. Gen terkait penyakit yang diketahui dilaporkan ke database publik, seperti Online Mendelian Inheritance in Man (OMIM) dan Genetic Association Database [1], mewakili sebagian kecil dari hubungan yang sebenarnya. Oleh karena itu, penemuan gen penyakit tetap penting. Pendekatan pemetaan gen tradisional melibatkan analisis keterkaitan dan studi asosiasi genom-lebar (GWAS) [2]. Karena jumlah persilangan yang terbatas dalam sampel keluarga, analisis keterkaitan biasanya hanya mengidentifikasi interval kromosom yang mungkin berisi hingga ratusan kandidat gen. Studi asosiasi seluruh genom juga dapat menghasilkan banyak wilayah yang masih harus diperiksa dalam karya mendatang. Validasi eksperimental dari begitu banyak gen kandidat memakan waktu dan mahal.

Karena mengintegrasikan beberapa sumber tambahan data penting untuk sifat penyakit gen, serangkaian alternatif komputasi berbasis jaringan telah diusulkan dalam dekade terakhir [3], [4], [5], [6]. Motivasi umum dari metode ini adalah bahwa gen yang menyebabkan penyakit yang sama atau serupa cenderung berkorelasi erat satu sama lain di jaringan biologis. Bukti khas bahwa model-model ini dapat dikategorikan adalah sebagai berikut [7]: penambangan teks literatur biomedis, penjelasan fungsional, jalur dan ontologi, hubungan fenotipe, hubungan gen intrinsik, data urutan, interaksi protein-protein, informasi regulasi, ortologis hubungan, dan informasi ekspresi gen. Misalnya, [8] menggunakan pendekatan penambangan teks untuk mengklasifikasikan fenotipe manusia berskala besar yang terdapat dalam database OMIM. Untuk menentukan kesamaan dalam jaringan interaksi protein-protein, [9] menyimpulkan koneksi gen-penyakit dengan menggunakan pengukuran jarak jaringan global yang disebut analisis lari dom walk. Selain itu, baru-baru ini [6] melakukan CATAPULT dan Katz pada jaringan heterogen, yang mengembangkan perluasan jalan acak menggunakan hitungan berjalan untuk menghitung kesamaan antara dua objek dan memasukkan informasi yang relevan dari spesies lain, seperti lalat buah dan tikus. Kelemahan utama dari metode berbasis jaringan tersebut adalah bahwa metode tersebut gagal dalam menggeneralisasi penyakit baru, yang belum ada studi keterkaitan gennya. Dalam hal ini, [10] membangun metode yang dihasilkan, Inductive Matrix Com pletion (IMC), berdasarkan sumber biologis yang berbeda, yang dapat diterapkan pada penyakit yang tidak terlihat pada waktu

• *Xiangxiang Zeng, Yinglai Lin, Yuying He dan Xiaoping Min (penulis yang sesuai) bekerja di Departemen Ilmu Komputer, Universitas Xiamen , Xiamen, Cina, 361005. (Xiangxiang Zeng, Yinglai Lin, Yuying Dia adalah penulis pertama.)*
• *Linyuan L ¨u bekerja di Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, dan Alibaba Research Center untuk Ilmu Kompleksitas, Alibaba Business College, Hangzhou Normal University, Hangzhou 311121, Cina.*
• *Xiangxiang Zeng dan Alfonso Rodr´ıguez-Pat´on bersama Departamento de Inteligencia Artificial, Universidad Polit´ecnica de Madrid, Campus de Montegancedo, Boadilla del Monte 28660, Madrid, Spanyol.*

tanggal diterima:??

pelatihan. Dari metode di atas untuk memprioritaskan gen yang relevan dengan penyakit tertentu, IMC standar berkinerja terbaik, meskipun mungkin menghasilkan pemahaman yang dangkal tentang fitur-fiturnya.

Kemajuan leverage pembelajaran mendalam diterapkan secara luas pada tugas klasifikasi dan representasi pembelajaran. Sistem yang dalam dengan banyak lapisan ini telah terbukti menghasilkan kinerja yang menjanjikan dalam mengekstraksi fitur data yang lebih kuat seperti [11], [12]. Proyek ini secara langsung menggunakan jaringan saraf konvolusional (CNN) [11], atau jaringan keyakinan dalam (DBN) [12], untuk mendapatkan faktor laten

2 PRELIMINARIES

Sebelum menjelaskan pendekatan kami, kami membahaspreliminar iessebagai berikut:

2.1 Stacked denoising Autoencoders

Sebuah autoencoder (AE) [17] adalah jaringan neural feedforward dengan satu lapisan tersembunyi, yang bertujuan untuk mempelajariterkompresi dan representasiterdistribusi secara otomatis dalamencoding fasedan merekonstruksi input bersihnya dalamdecoding fase. Diberikan masukan, x, model menerapkan encoder, $f(\cdot)$, untuk memetakannya ke fitur tersembunyi $f(x)$, sedangkan decoder $g(\cdot)$ memetakan fitur tersembunyi ke lapisan keluaran $\hat{x}$ sebagai versi rekonstruksi dari x,sehingga $g(f(x)) = x$. Parameterautoencoder yang dipelajari untuk meminimalkankesalahan rekonstruksi, diukur dengan beberapa fungsi kerugian $L(x, g(f(x)))$ kami.

Gambar 1:. Ilustrasi Stacked denoising Autoencoders.

untuk informasi konten. Referensi [13] dan [14] menggunakan Bayesian stacked denoising autoencoders (SDAE), dan SDAE terpinggirkan, untuk penyaringan kolaboratif. Dalam makalah ini, kami mengusulkan model yang fleksibel dan kuat, yang bersama-sama menggunakan faktor laten yang dalam dari gen dan pembelajaran penyakit dari informasi samping dan melakukan penyaringan kolaboratif dari matriks asosiasi gen-penyakit. Selanjutnya, upaya dilakukan untuk mengubah pengoptimalan. Untuk fenotipe tertentu, sulit untuk memverifikasi bahwa bukti yang tidak teramati menunjukkan apakah hubungan semacam itu tidak ada atau tersembunyi. Terinspirasi oleh berbagai metode pembelajaran yang berfokus pada pembelajaran PU untuk penyelesaian matriks dalam literatur sebelumnya [15], [16], kami menggunakan pembelajaran PU dalam pekerjaan kami. Eksperimen menunjukkan bahwa model hibrida kami jauh mengungguli keadaan seni. Secara khusus, kontribusi utama makalah ini adalah sebagai berikut:

Kami merangkum kontribusi utama kami sebagai berikut. Pertama, kami mengusulkan model Deep Collaborative Filtering (DCF), yang mengintegrasikan pembelajaran representasi mendalam dan penyelesaian matriks. Kedua, kami menghadirkan Stacked Denoising Autoen coders (SDAE) sebagai algoritma pembelajaran mendalam untuk mengurangi dimensi menjadi ukuran gen yang dapat dikelola dengan formasi yang relevan. Selain itu, SDAE tidak hanya mengatasi kebisingan dan sifat lengkap dari berbagai sumber data, tetapi juga mencapai tingkat kompresi yang lebih tinggi yang membuatnya lebih sesuai untuk kasus biologis. Selain itu, pembelajaran PU diadopsi dalam memproses matriks asosiasi. Ketiga, kami membandingkan model kami dengan metode penyakit OMIM yang canggih. Hasil eksperimen menunjukkan keunggulan DCF dalam mengambil hubungan yang sebenarnya dan pengukuran presisi-recall. Kami juga menganalisis kebaruan prediksi untuk penyakit baru dan asosiasi baru. Dengan menutupi fitur parsial, kami mengeksplorasi pentingnya fitur dari fitur gen yang diberikan pada

Untuk memaksa autoencoder menghasilkan fitur yang lebih kuat daripada identitas sederhana, autoencoder denoising (DAE) diusulkan oleh [18]. Alih-alih langsung menggunakan input asli, x , DAE menggabungkan beberapa kriteria kerusakan yang berbeda ke dalam data mentah, termasuk noise masking atau noise Gaussian isotropik aditif. Proses penambahan kebisingan meningkatkan kekokohan fitur yang terlatih dan secara berurutan pada model yang dihasilkan. Dalam studi ini, kami mengadopsi Stacked DAE (SDAE) yang diusulkan oleh [19], yang menumpuk DAE untuk membentuk jaringan dalam, memungkinkan SDAE untuk mempelajari fitur hierarki dari lapisan input ke lapisan tersembunyi. Model A SDAE meminimalkan masalah regularized dan menangani kesalahan reconstruc tion dalam analog cara untuk AE, didefinisikan sebagai berikut:

$$\min_{\{W_l\},\{b_l\}} \|x - \hat{x}\|_F^2 + \lambda \sum_l \|W_l\|_F^2 \quad (1)$$

dimana L adalah jumlah lapisan, $W_l$adalah matriks bobot dan b_l adalah vektor bias dari lapisan $l \in \{1, ..., L\}$, yang dapat dipelajari dengan algoritma propagasi balik. λ adalah parameter regularisasi dan $\|\cdot\|_F$menunjukkan norma Frobenius.

Lapisanpertama $L/2$ dari model bertindak sebagai pembuat enkode, dan lapisanterakhir $L/2$ bertindak sebagai dekoder. Lapisan tengah adalah kunci yang memungkinkan SDAE mengurangi dimensi dan

mengekstrak representasi informasi samping yang efektif.

2.2 Faktorisasi Matriks

Pada bagian ini, pertama kali kami perkenalkan penyelesaian matriks rank rendah dan penyelesaian matriks induktif. Penyelesaian matriks peringkat rendah kerangka yang diusulkan. tinjauan SDAE yang disederhanakan diberikan terlebih dahulu, diikuti oleh

Makalah ini disusun sebagai berikut: Pada Bagian 2,

diskusi tentang keterbatasan teknik penyelesaian matriks tradisional. Di Bagian 3, penjelasan rinci tentang pendekatan yang kami usulkan diberikan. Kumpulan data percobaan, metrik evaluasi dan analisis hasil disajikan di Bagian 4. Kesimpulan disajikan di Bagian 5. di

(MC) adalah salah satu metode penyaringan kolaboratif yang paling populer dan sukses untuk sistem pemberi rekomendasi [20]. Tujuannya adalah untuk memulihkan matriks yang mendasari dengan menggunakan entitas yang diamati dari P_{ij}, yang didefinisikan sebagai berikut:

X

$$\min_{W,H} \sum_{(i,j)\in\Omega} (P_{ij} - (WH^T)_{ij})^2 + \frac{\lambda}{2}(kWk^2_F + kHk^2_F) \quad (2)$$

mana λ adalah parameter regularisasi dan $\Omega \in N_x \times N_y$ adalah entri yang diamati dari matriks dasar yang sebenarnya. Dengan asumsi matriks dimodelkan menjadi rank rendah, yang berarti $W = R^{N_x \times k}$ dan $H = R^{N_y \times k}$ berbagi ruang laten berdimensi rendah, memenuhi $k\ N_x$, N_y.

2.3 Penyelesaian Matriks Induktif Penyelesaian

matriks tradisional didasarkan pada pengaturan transduktif. Selain itu, semua pendekatan penyelesaian matriks mengalami ketersebaran yang ekstrim dari matriks yang diamati dan masalah cold-start. Untuk mengurangi batasan ini, penyelesaian matriks induktif (IMC) [21] dikembangkan, yang dapat diinterpretasikan sebagai generalisasi dari formulasi multi-label transduktif dan memungkinkan kami untuk menggabungkan informasi samping. Teknologi ini diterapkan untuk membuat prediksi tentang asosiasi penyakit gen [10]. Secara khusus, informasi samping dari kedua entitas diberikan dalam dua matriks: $x_i \in R^{N_g}$ menunjukkan vektor fitur untuk gen i dan $y_j \in R^{N_d}$ menunjukkan vektor fitur untuk penyakit j. IMC mengasumsikan bahwa matriks asosiasi yang mendasari dihasilkan dengan menerapkan vektor fitur gen dan penyakit ke matriks peringkat rendah, yang dipelajari dari serangkaian pelatihan asosiasi gen-penyakit, yang dirumuskan sebagai:

$$\min_{W,H} \sum_{(i,j)\in\Omega} `(P_{ij}, x^T_i WH^T y_j) + \frac{\lambda}{2}(kWk^2_F + kHk^2_F) \quad (3)$$

di mana `*adalah* fungsi kerugian yang mengukur deviasi antara prediksi dan pengamatan, misalnya, kehilangan *kuadrat*`$_{persegi}$(a,b)=(a-b)^{2}dan hilangnya logistik `$_{log}$(a,b)= log (1+e^{-ab}).

2.4 Pembelajaran PU

Metode pembelajaran mesin yang ada biasanya memperlakukan gen penyakit yang diketahui sebagai positif dan gen yang tidak diketahui sebagai negatif, ini dapat menyebabkan hukuman yang berlebihan pada sampel yang tidak berlabel karena mungkin ada gen penyakit yang tidak diketahui dalam sampel yang tidak berlabel. Untuk menghindari hal ini, kami memperlakukan sampel yang tidak diketahui sebagai tidak berlabel, dan melakukan pembelajaran Positive-Unlabeled (PU) di set. Jadi, masalah optimasi berikut dianggap [15]:

Misalkan $X = x_1, ..., x_{Ng}^T$ menunjukkan matriks fitur terlatih

3.1 Kerangka Umum

Dengan intuisi pengurangan dimensi , kami merangkul SDAE dan Analisis Komponen Utama (PCA) untuk mendapatkan fitur berdimensi rendah dari informasi gen dan sisi penyakitdigambar oleh himpunan X dan himpunan Y. , masing-masingDiberikan matriks asosiasi penyakit gen P, DCF secara bersama-sama memperoleh fitur terkompresi dari informasi samping dan mempelajari matriks proyeksi $Z = WH^T$ menggunakan entri yang diamati dari P *di* bawah pengaturan pembelajaran PU, diformulasikan sebagai:

$$L^+(P, XZY^T) + \alpha L^-(P, XZY^T) + \frac{\lambda}{2}(kWk^2_F + kHk^2_F) \quad (5)$$ di mana $L^+(W, H)$ adalah fungsi kerugian yang sesuai dengan entri positif, dan $L^-(W, H)$ adalah hilangnya sampel negatif. Karena penyelesaian matriks induktif standar biasanya mengabaikan entri negatif, untuk mengatasi masalah ini, nilai kecil α bias estimasi dari contoh yang tidak teramati

menuju nol.

3,2 Jauh Collaborative Filtering

dari gen, di mana *i*baris ke-adalah gen *i* vektor fitur $x_i \in R^{f_g}$, $i = 1, ..., N_g$, dan misalkan $Y = [y_1, ..., y_N d]^T$menyatakan fitur dilatih matriks penyakit, dimana *j*baris -th adalah penyakit *j* vektor fitur $y_j \in R^{f_d}$, $j = 1,..., N_d$.Secara khusus, SDAE dan PCA ditetapkan untuk ruang dimensi yang sama *m* pada fitur laten dari setiap kumpulan data terkait. Misalnya, dengan mempertimbangkan jaringan fungsional gen, $G \in R^{N_g \times N^g}$, dan kami menggunakan $\bar{G}$ untuk menunjukkan keluaran yang direkonstruksi. Sebagaimana dibahas sebelumnya, fungsi tujuan SDAE bekerja untuk meminimalkan kerugian antara $\bar{G}$ dan G untuk menghubungkanbersih inputdengan fitur laten. Sekarang, $N_g \times m$ mewakili

min $_{W,H}$ X $_{i,j:P_{ij}=1}$ `$(P_{ij}, WH^T)$ + $\alpha$ X $_{i,j:P_{ij}=0}$ `(P_{ij}, WHT) ukuran representasi vektor dimensi rendah melalui SDAE, yangdimensi m jauh lebih rendah dari G.Dalam makalah ini, kami melakukan SDAE pada data ekspresi gen,

$+ \frac{\lambda}{2}(kWk^2_F + kWk^2_F)$ (4) di mana α menunjukkan bobot fungsi kerugian sampel yang tidak berlabel.

3 METODE

Meskipun IMC mampu menghasilkan keakuratan yang mengesankan dan mengintegrasikan bukti yang kaya untuk memprediksi hubungan gen-penyakit, masih ada sesuatu yang diinginkan, terutama dalam hal data biologis yang berkualitas rendah dan berdimensi tinggi. Di sini, kami menyajikan model kami, Deep Collaborative Filtering (DCF). Metode ini didasarkan pada pengakuan berikut: literatur yang ada menunjukkan bahwa SDAE adalah cara yang ampuh untuk membangun representasi berkualitas tinggi tanpa memerlukan pengetahuan sebelumnya. Selain itu, kami mempertimbangkanbioinfor

interaksi fungsionalgen dan asosiasi gen-fenotipe dari data matics, yang terdiri dari pasangan gen-penyakit positif (menunjukkan bahwa suatu gen terkait dengan penyakit tertentu), tetapi tidak ada pasangan negatif yang tersedia, yaitu gen yang tidak terkait dengan penyakit. Akibatnya, masalah ini secara alami dipelajari dalam kerangka pembelajaran PU, di mana entri yang diamati dan yang dilayani diberi hukuman berbeda dalam tujuan. Ke pada akhirnya, kombinasi tugas pembelajaran SDAE dan PU yang kami usulkan bermanfaat untuk melatih model yang lebih ekspresif. Penjelasan singkat tentang kerangka DCF diberikan di bawah ini.

di mana himpunan Ω mencakup entri positif dan negatif, sehingga $\Omega = \Omega^+ \ S\Omega^-$, misalkan Ω^+ menunjukkan sam yang diamati dan Ω^- menunjukkan entri yang hilang dipilih sebagai negatif. Untuk penyelesaian matriks induktif bias, nilai α adalah

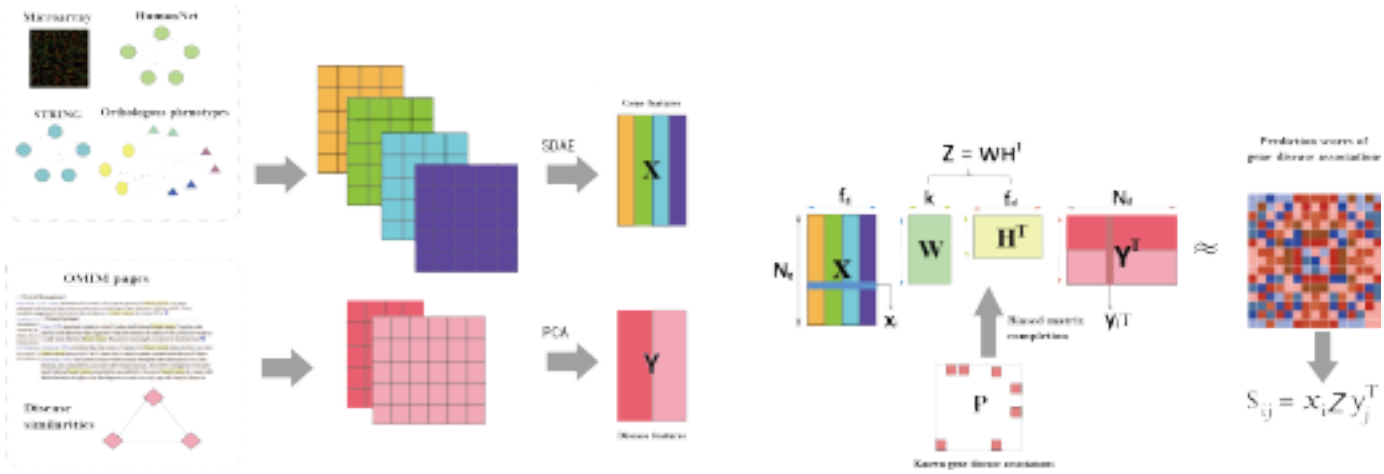
parameter kunci sehingga $\alpha < 1$ karena bobot penalti untuk entri yang diamati harus lebih besar dari yang hilang.

spesies lain. Secara khusus, matriks dokumen istilah dari halaman web OMIM dan kesamaan penyakit mempertahankan analisis PCA untuk ketersebaran yang ekstrim. Gambaran umum DCF disajikan pada Gambar 2.

Matriks asosiasi gen-penyakit diberikan sebagai $P \in R^{N_g \times N^d}$, di mana N_g adalah jumlah gen dan N_d adalah jumlah penyakit. Ketika $P_{ij} = 1$, menyimpulkan gen i terkait dengan penyakit j sedangkan nol menunjukkan hubungan tersebut tidak teramati. Setelah proses ekstraksi ciri, kita membangun fungsi dekomposisi untuk memulihkan matriks peringkat rendah $Z \in R^{f_g \times f^d}$ dari matriks asosiasidiketahui P *yang* dengan bentuk $Z = WH^T$, di mana $W \in R^{f_g \times k}$dan $H \in R^{f_d \times k}$, k N_g, N_d. Masalah pengoptimalan model kami diparameterisasi sebagai:

$$\min_{W,H} \sum_{(i,j) \in \Omega^+} (P_{ij} - x_i W H^T y^T_j)^2 + \sum_{(i,j) \in \Omega^-} (P_{ij} - x_i W H^T y^T_j)^2 + \frac{\lambda}{2}(kWk^2_F + kHk^2_F) \quad (6)$$

IEEE 1545-5963 (c) 2018. Penggunaan pribadi diperbolehkan, tetapi publikasi ulang / redistribusi memerlukan izin IEEE. Lihat http://www.ieee.org/publications_standards/publications/rights/index.html untuk informasi lebih lanjut.
Artikel ini telah diterima untuk diterbitkan pada terbitan jurnal ini yang akan datang, tetapi belum diedit sepenuhnya. Konten dapat berubah sebelum publikasi akhir.
Informasi kutipan: DOI 10.1109 / TCBB.2019.2907536, Transaksi IEEE / ACM pada Biologi Komputasi dan Bioinformatika



Gbr. 2: Alur pendekatan yang diusulkan. DCF pertama-tama memetakan berbagai gen dan sumber terkait penyakit ke ruang berdimensi rendah dan menyusun representasi informatif menggunakan algoritme reduksi dimensionalitas SDAE dan PCA. Kemudian menerapkan skema bias pembelajaran PU untuk melakukan penyelesaian matriks pada asosiasi penyakit gen yang diketahui.

Dalam percobaan kami, nilai bias α dan parameter regulasi ter λ dipilih melalui pencarian grid.

Kemudian, kami memperkirakan kemungkinan skor

interaksi berpasangan antara gen *i* dan penyakit *j* sebagai:

$$skor(i, j) = x_i WH^T y^T_j \quad (7)$$

di mana skor yang lebih tinggi menunjukkan kemungkinan yang lebih tinggi bahwa gen *i* berkorelasi dengan penyakit *j*.

3.3 Optimasi Optimasi

fungsi tujuan menjadi konveks jika kita mengubah satu variabel dan memperbaiki variabel lainnya. Untuk melakukannya, kami memperbaiki *W* dan memperbarui *H* (dan sebaliknya) untuk mengoptimalkan (6). Dua pendekatan utama untuk memilih entri negatif adalah: (1) Subsampel himpunan dengan ukuran yang mirip dengan entri positif yang diamati (metode "Sub-sampel") atau (2) memperlakukan semua entri yang hilang sebagai negatif (metode "Penuh"). Umumnya, pendekatan Penuh lebih baik daripada pendekatan Sub-sampel, tetapi sering kali tidak efisien karena sejumlah besar sampel negatif. Baru-baru ini, [22] mengembangkan metode optimasi yang efisien untuk metode penuh, dengan berurutan mengoptimalkan sepasang kolom dari *W* dan *H* daripada mengoptimalkan semua nilai dalam *W* dan *H*.Jadi, kami mengadopsi metode mereka dan memperluasnya ke pembelajaran PU. Untuk menyelesaikan kira-kira (6), kami mengeksploitasi penurunan gradien konjugasi dengan kerugian kuadrat dalam iterasi.

4 Results dan Diskusi

Pada bagian ini, kami mengevaluasi kinerja model DCF kami dengan lima dataset dunia nyata dalam biologi dan membandingkan DCF kami dengan algoritma mutakhir.

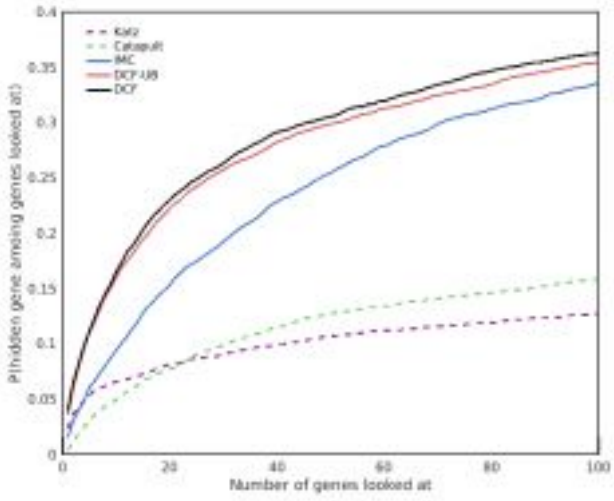
4.1 Dataset

- **Data tambahan untuk ekstraksi fitur gen**
  - **microarray** Pengukurandari tingkat ekspresi gen dalam sampel jaringan yang berbeda, diperoleh dari BioGPS (www.biogps.org) dan Connectivity Map (www.broadinstitute.org/cmap) oleh [10]. Rincian microarray dan gen / penyakit lainnya ditunjukkan pada Tabel1.
  - **HumanNet** adalah jaringan gen fungsional skala besar yang menggabungkan kumpulan data yang beragam, termasuk ekspresi mRNA, interaksi protein-protein, data kompleks protein, dan genomik komparatif (tetapi bukan data penyakit atau fenotipe) [23].
  - **Asosiasi gen-fenotipe** dari spesies lain yang relatif lebih kaya dibandingkan dengan studi gen-penyakit pada manusia. Kami menggunakan fenotipe untuk delapan spesies berbeda: tanaman [24], cacing [25], [26], lalat buah [27], tikus [28], ragi [29], [30], [31], [32] , Escherichia coli [33], ikan zebra [34] dan ayam [35]. Rinciannya terangkum dalam Tabel 2.
  - **STRING** [36] merupakan jaring interaksi gen-gen dengan ukuran besar, diunduh dari NDEx (http: // www.ndexbio.org), diproses oleh [37], yang secara kuantitatif menggabungkan beberapa studi dan jenis antar tindakan dan menyesuaikan bobot untuk setiap pasangan gen pada bobot total bukti.
- **Data tambahan untuk ekstraksi ciri penyakit -**
  Jaringan kesamaan penyakit MimMiner [8] memberikan jaringan berbobot antar penyakit.
  - **Halaman web untuk penyakit OMIM yang** digunakan di [10], termasuk bagian 'Gambaran Klinis' dan 'Man agement Klinis' dari halaman web yang mendokumentasikan gejala, pengobatan dan tanggapan oleh pasien, dan studi terkait tentang efek dari kursus yang berbeda terapi.
- *hubungan* **gen-penyakit yang diketahui**
  - **Asosiasi penyakit gen** terkandung sebagai data evaluasi, diunduh dari proyek Online Mendelian Inher itance in Man (OMIM) pada tanggal 11 Agustus 2011 yang digunakan di [6]. Fenotipe OMIM telah menjadi kumpulan data standar untuk evaluasi prediksi asosiasi penyakit gen [3], [4], [5], [6], [9], [10], [38], [39].

indeks gen / nama penyakit # gen / penyakit

1
Microarray 8,755 2 HumanNet 18,714
3 STRING 18,266
4 Kemiripan penyakit 3.209
5 OMIM 3.209

TABEL 1: Detail gen / penyakit. The ta
menunjukkan ble ukuran gen / penyakit yang digunakan dalamkami
ukuran experiments.The Penyakit similari
ikatan danOMIM jauh lebih besar dari 3209, tapi
kami hanya peduli tentang beberapa penyakit.



Nama Spesies	Indeks	#Phenotypes	#Associations
1	Manusia	3209	3954
2	Tanaman	1.137	12.010
3	Worm	744	30.519
4	Buah Fly	2503	68.525
5	ikan zebra	1143	4500
6	Escherichia coli	324	72.846
7	Chicken	1.188	22.150
8	Tikus	4662	75.199
9	Ragi	1.243	73.284

TABEL 2: Detil untuk jenis. Tabel menunjukkan ukuran spesies yang digunakan dalam percobaan kami.

4.2 Metrik evaluasi

Mirip dengan ProDiGe [38], CATAPULT [6] dan IMC [10], sebagai ukuran untuk membandingkan kinerja metode yang berbeda, kami menggunakan fungsi distribusi kumulatif (CDF) dari peringkat (yaitu, probabilitas bahwa peringkat di mana pasangan gen-penyakit yang tersembunyi diambil kurang dari ambang batas *r*). Di sini, kami melaporkan hasil untuk $r \leq 100$.

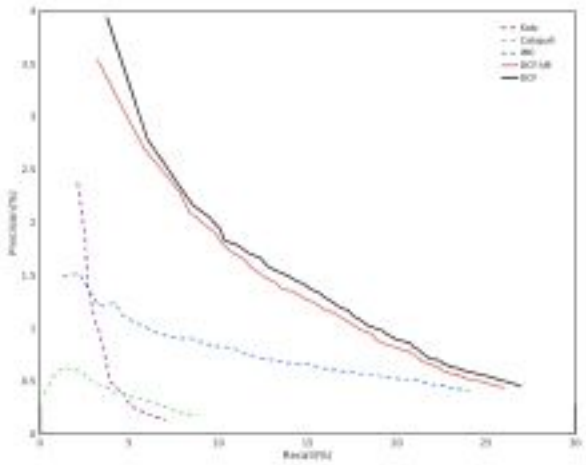
Kami juga mengevaluasi dan membandingkan berbagai metode menggunakan ukuran presisi dan perolehan yang lebih dikenal. Presisi mengukur fraksi positif sejati (gen) pulih dalamtop.*r* prediksi Ingat adalah rasio positif sejati pulih dalamtop*r* prediksiuntuk suatu sifat untuk jumlah positif sejati di set tersembunyi. Disajikan dalam Gambar 3 (b) adalah plot tingkat presisi-recall untuk nilai ambang yang berbeda, *r* , dalam rentang $0 \leq r \leq 100$.

4.3 Dasar dan Pengaturan Eksperimental

Hasil prediksi didasarkan pada perbandingan rinci antara model yang terdaftar sebagai berikut:

1) *Katz:* Pengukuran Katz [6] adalah metode berbasis grafik untuk menemukan kemiripan node dengan node tertentu dengan menghitung berapa banyak jalur dari panjang yang berbeda ada di antara pasangan.
2) *CATAPULT:* Menyisir dATa Lintas spesies menggunakan Teknik Pembelajaran Positif Tanpa Label [6] adalah metode pembelajaran terbimbing yang melatih pengklasifikasi mesin vektor dukungan bagging melalui jaringan yang heterogen.
3) *IMC:* Inductive Matrix Completion [10] adalah pendekatan baru yang dikembangkan oleh (Jain dan Dhillon, 2013) yang menggabungkan beberapa fitur saat mencoba mempelajari asosiasi gen-penyakit.
4) *DCF-UB:* Pemfilteran Kolaborasi Mendalam yang Tidak Memihak menggunakan SDAE untuk mengelola fitur gen. Proses penyelesaian matriks tetap sama seperti di IMC.

(a) CDF peringkat



(b) Kurva Precision-Recall

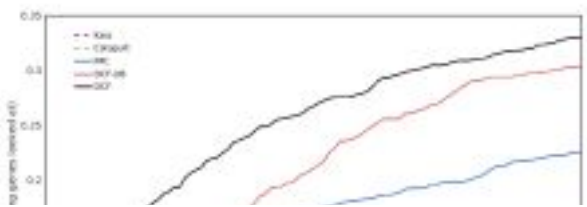
Gbr. 3: Perbandingan kinerja dengan metode mutakhir lainnya untuk prioritas gen penyakit. (a) Semua hasil dirangkum dalam validasi silang tiga kali lipat untuk menggambarkan probabilitas bahwa asosiasi gen sejati diambil di 100 teratas. (b) Performa setiap metode dinilai oleh area di bawah kurva presisi-recall.

5) *DCF:* Pendekatan kami diusulkan seperti dijelaskan di atas. Ini adalah model hybrid, yang menyatukan struktur pembelajaran mendalam, SDAE, dengan penyelesaian matriks induktif PU.

Untuk set data Microarray, pertama-tama kita mengubah rentang ekspresi setiap gen secara linier menjadi antara nol dan satu, lalu membangun SDAE 4 lapis. Jumlah unit tersembunyi mencapai '1000-500-200-100'. Masking noise diproses dengan tingkat kebisingan 0,1 untuk mendapatkan versi input yang rusak, di mana elemen untuk setiap sampel disetel ke nol secara acak pada tingkat kerusakan yang telah ditetapkan. Kami mengadopsi sigmoid untuk aktivasi dan dropout untuk generalisasi setiap lapisan kecuali lapisan terakhir, menetapkan kecepatan pembelajaran menjadi 0,1 dan epoch menjadi 500.

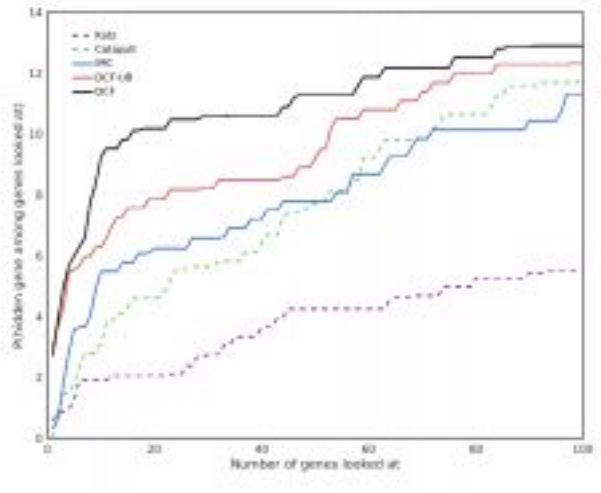
Untuk fenotipe ortolog dan jaringan interaksi gen-gen, kami melatih arsitektur mendalam pada kombinasi

Artikel ini telah diterima untuk diterbitkan pada terbitan jurnal ini yang akan datang, tetapi belum diedit sepenuhnya. Konten dapat berubah sebelum publikasi akhir.
Informasi kutipan: DOI 10.1109 / TCBB.2019.2907536, Transaksi IEEE / ACM pada Biologi Komputasi dan Bioinformatika



metrik evaluasi. Hasil validasi silang tiga kali lipat menunjukkan bahwa DCF memiliki peluang hampir 37% untuk mendapatkan gen sejati dalam 100 prediksi teratas untuk penyakit; Padahal, IMC hanya memiliki sekitar 33%. Kami menerapkan CATAPULT dan

(a) penyakit Singleton



(b) Asosiasi yang baru ditemukan

Gambar. 4: Evaluasi pada penyakit tunggal dan asosiasi yang baru-baru ini dilaporkan untuk menunjukkan peringkat distribusi kumulatif dan ingatan, masing-masing. (a) Kinerja pada penyakit tunggal. (b) kinerja pada 493 asosiasi yang baru ditemukan.

dari parameter berikut: jaringan 2-lapisan dari node '200-100', tingkat kerusakan 0,3, dan kecepatan pembelajaran 0,1. Untuk mempercepat proses konvergensi, kami melatih DAE dalam beberapa batch sampel. Ukuran tumpukan nomor disetel ke 128. Matriks bobot dan vektor bias dioptimalkan melalui penurunan gradien stokastik dan diperbarui di setiap kelompok.

Selanjutnya, dimensionalitas dari setiap faktor laten yang dipelajari untuk gen dan penyakit disetel ke 100; kemudian, peringkat matriks peringkat rendah, k, diatur ke 200, sesuai dengan itu. Untuk menguji metode kami, kami melakukan validasi silang tiga kali lipat. Semua pasangan gen-penyakit secara acak dibagi menjadi tiga set yang ukurannya kira-kira sama. Setelah pencarian grid, kami menemukan bahwa DCF berkinerja terbaik ketika parameter regularisasi λ = 0,01 dan α = 0,05 (α menentukan penalti entri yang tidak teramati menuju nol).

4.4 Kinerja Keseluruhan

Seperti yang ditunjukkan pada Gambar 3, dari metode kompetitif, model kami mencapai peningkatan yang berbeda dalam dua kekuatan prediksi. Sepengetahuan kami, data ekspresi gen Katz pada jaringan gabungan, yang berisigen-gen jaringan interaksidari HumanNet dan jaringan gen-fenotipe dari 8 spesies. Untuk CATAPULT, 16% gen tersembunyi berada di peringkat 100 teratas; sedangkan Katz mempertahankan 12%, yang mungkin dikaitkan dengan tidak adanya fitur penyakit.

Gambar 3 (b) menyajikan hasil presisi-recall yang IMC, CmembandingkanATAPULT, Katz dan model kami. Outper DCF membentuk IMC sehubungan denganstandar *presisi*@ 100 dan *recall*@ 100 pengukuran, yang menunjukkan kekuatan vektor laten yang dipelajari oleh arsitektur dalam danvali tanggal, seperti yang diharapkan, bahwa SDAE memainkan peran kunci dalam improv sangat berisik dan berkorelasi. Dibandingkan dengan metode kompresi fitur yang sering digunakan, PCA, hasil menunjukkan bahwa SDAE menangani lebih baik tantangan komputasi yang timbul dari data biologis yang tidak lengkap dan berisik.

Selain itu, DCF-UB merupakan metode mendalam tanpa pembelajaran PU dalam fase kolaboratif filtering. Gambar 3 menunjukkan bahwa strategi bias, yang digunakan dalam model DCF, bekerja untuk mendapatkan hasil yang lebih baik dalam tugas identifikasi gen-penyakit.

4.5 Induksi penyakit baru

Kami juga menggarisbawahi kekuatan penemuan penyakit baru dari metode DCF yang diusulkan. Penyakit dengan hanya satu asosiasi yang diketahui dalam kumpulan data, tetapi tidak ada dalam pelatihan (penyakit tunggal), dapat menjelaskan kapasitas untuk mengekstrak fitur efektif dari informasi samping. Gambar 4 (a) memberikan perbandingan kinerja untuk tugas ini. Peringkat CDF untuk DCF menggambarkan keunggulan untuk memprediksi 100 gen teratas yang terkait dengan penyakit baru.

4.6 Evaluasi pada asosiasi yang baru ditemukan

Karena validasi silang pada data retrospektif mungkin mengarah pada hasil yang terlalu optimis, validasi silang tidak sesuai untuk masalah ini. Untuk mengevaluasi kemampuan model untuk memprediksi asosiasi yang baru ditemukan, kami melatih semua metode menggunakan 12.331 gen yang terkait dengan 3.209 penyakit OMIM yang dikumpulkan hingga Mei 2007 [40] dan menjalankan pengujian pada asosiasi baru yang muncul dari Mei 2007 hingga Juli 2014. Edisi baru be tween Mei 2007 dan Juli 2014, membawa 493 asosiasi yang baru ditemukan dengan 242 gen yang tidak terkait sebelumnya. Dengan demikian, eksperimen ini juga mengevaluasi kinerja gen baru yang direkomendasikan. Kinerja semua metode dalam validasi ini ditampilkan pada Gambar 4 (b). Kita dapat melihat bahwa DCF mencapai kinerja terbaik dan DCF UB berkinerja lebih baik dari IMC. Selain itu, dibandingkan dengan hasil pada Gambar 3 (a), penurunan nyata dalam kinerja semua metode diamati, karena asosiasi yang lebih sedikit diketahui pada set pelatihan.

4.7 Investigasi tentang pentingnya fitur untuk gen Pada subbagian ini, kami mempelajari pentingnya 4 jenis fitur gen, dengan mengganti beberapa fitur secara acak dengan matriks full-zero dengan ukuran yang sama. Seperti yang bisa kita lihat pada Gambar

5, STRING adalah sinyal terpenting untuk prediksi gen penyakit, dan ortologis adalah yang berikutnya. Selain itu, lebih banyak fitur gen meningkatkan kinerja.

Akibatnya, dengan menggabungkan model SDAE secara mulus untuk informasi sisi tambahan dan filter kolaboratif untuk matriks asosiasi gen-penyakit, model kami mempelajari representasi yang jauh lebih bermakna untuk setiap gen dan penyakit dan memberikan prediksi yang lebih akurat.

5 Conclusion

In this paper, we presented a deep collaborative filtering model that bridges SDAE and matrix completion. This method also offers a scalable platform for integrating var ious categories of features of both genes and diseases. Quantitative results reveal our method, by taking advan tage of deep architectures, outperforms other state-of-the-art baselines. Regarding further work, the squared loss used in the DCF is unstable with the noise of the dataset. The score matrix entries equal zero or one, so there might be a more suitable loss function for the binary matrix.

In this study, we simply combine features obtained in dependently from microarray, gene associations networks and gene-phenotype network. However, this combination may ignore the intrinsic relationship among features that are shared with different auxiliary data. For future works, deep canonical correlation analysis [41] may be a promising method to extract features that are shared by different types of auxiliary data (views). On the other hand, other deep learning models, such as convolutional neural networks, could be unified in our DCF framework to further boost performance.

Acknowledgments

The work was supported by the National Natural Sci ence Foundation of China (Grant Nos. 61472333, 61772441, 61872309, 61472335, 61272152, and 41476118), Natural Sci ence Foundation of the Higher Education Institutions of Fujian Province (No. JZ160400), President Fund of Xiamen University (No. 20720170054). X. Zeng is supported by Juan de la Cierva position (code: IJCI-2015-26991). A. Rodr´ıguez Paton is partially supported by Spanish TIN2016-81079-R, ´ (MINECO AEI/FEDER, EU) and Madrid Gov. B2017/BMD 3691, InGEMICS-CM(FSE/FEDER, EU).

References

[1] KG Becker, KC Barnes, TJ Bright, and SA Wang, “The genetic association database.” *Nature Genetics*, vol. 36, tidak. 5, pp. 431–432, 2004.

[2] D. Botstein and N. Risch, “Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease.” *Nature Genetics*, vol. 33, tidak. 33 Suppl, pp. 228–237, 2003.

[3] X. Wu, R. Jiang, MQ Zhang, and S. Li, “Network-based global in ference of human disease genes,” *Molecular systems biology*, vol. 4, tidak. 1, hal. 189, 2008.

[4] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, “As sociating genes and protein complexes with disease via network propagation.” *Plos Computational Biology*, vol. 6, tidak. 1, hal. e1000641, 2010.

[5] Y. Li and JC Patra, “Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network,” *Bioinfor matics*, vol. 26, tidak. 9, pp. 1219–1224, 2010.

[6] UM Singh-Blom, N. Natarajan, A. Tewari, JO Woods, IS Dhillon, and EM Marcotte, “Prediction and validation of gene disease associations using methods inspired by social network analyses,” *PloS one*, vol. 8, tidak. 9, 2013.

[7] RM Piro and CF Di, “Computational approaches to disease gene prediction: rationale, classification and successes,” *Febs Jour nal*, vol. 279, no. 5, pp. 678–696, 2012.

[8] MAV Driel, J. Bruggeman, G. Vriend, GB Han, and JAM Leunissen, “A text-mining analysis of the human phenome,” *European Journal of Human Genetics*, vol. 14, tidak. 5, pp. 535–542, 2006.

[9] S. Kohler, S. Bauer, D. Horn, and PN Robinson, “Walking the in- ¨ teractome for prioritization of candidate disease genes,” *American Journal of Human Genetics*, vol. 82, tidak. 4, pp. 949–958, 2008.

[10] N. Natarajan and IS Dhillon, “Inductive matrix completion for predicting gene–disease associations,” *Bioinformatics*, vol. 30, tidak. 12, pp. i60–i68, 2014.

[11] S. Dieleman and B. Schrauwen, “Deep content-based music rec ommendation,” in *International Conference on Neural Information Processing Systems*, 2013, pp. 2643–2651.

[12] X. Wang and Y. Wang, “Improving content-based and hybrid music recommendation using deep learning,” in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 627–636.

[13] H. Wang, N. Wang, and D.-Y. Yeung, “Collaborative deep learning for recommender systems,” in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1235–1244.

[14] S. Li, J. Kawale, and Y. Fu, “Deep collaborative filtering via marginalized denoising auto-encoder,” in *ACM International on Conference on Information and Knowledge Management*, 2015, pp. 811– 820.

[15] C.-J. Hsieh, N. Natarajan, and IS Dhillon, “Pu learning for matrix completion.” in *ICML*, 2015, pp. 2445–2453.

[16] N. Natarajan, N. Rao, and I. Dhillon, “Pu matrix completion with graph information,” in *IEEE International Workshop on Computa tional Advances in Multi-Sensor Adaptive Processing*, 2016, pp. 37–40.

[17] B. Scholkopf, J. Platt, and T. Hofmann, “Greedy layer-wise training ¨ of deep networks,” in *International Conference on Neural Information Processing Systems*, 2006, pp. 153–160.

[18] P. Vincent, H. Larochelle, Y. Bengio, and PA Manzagol, “Extract ing and composing robust features with denoising autoencoders,” in *International Conference on Machine Learning*, 2008, pp. 1096–1103.

[19] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and PA Manzagol, “Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion.” *Journal of Machine Learning Research*, vol. 11, tidak. 12, pp. 3371–3408, 2010.

[20] C. Volinsky, C. Volinsky, and C. Volinsky, *Matrix Factorization Techniques for Recommender Systems*. IEEE Computer Society Press, 2009.

[21] P. Jain and IS Dhillon, “Provable inductive matrix completion,” *Computer Science*, 2013.

[22] H.-F. Yu, M. Bilenko, and C.-J. Lin, “Selection of negative samples for one-class matrix factorization,” in *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, 2017, pp. 363–371.

[23] I. Lee, UM Blom, PI Wang, JE Shim, and EM Marcotte, “Prioritizing candidate disease genes by network-based boosting of genome-wide association data,” *Genome Research*, vol. 21, tidak. 7, hal. 1109, 2011.

[24] S. David, W. Christopher, L. Philippe, TZ Berardini, GH Margarita, F. Hartmut, D. Li, M. Tom, M. Robert, and P. Larry, “The arabidopsis information resource (tair): gene structure and function annotation,” *Nucleic Acids Research*, vol. 36, tidak. Database issue, pp. D1009–D1014, 2008.

[25] N. Chen, TW Harris, I. Antoshechkin, C. Bastiani, T. Bieri, D. Blasiar, K. Bradnam, P. Canaran, J. Chan, and CK Chen, “Wormbase: a comprehensive data resource for caenorhabditis bi ology and genomics,” *Nucleic Acids Research*, vol. 33, tidak. Database issue, p. D383, 2005.

[26] RA Green, HL Kao, A. Audhya, S. Arur, JR Mayers, HN

Fridolfsson, M. Schulman, S. Schloissnig, S. Niessen, and K. La band, “A high-resolution c. elegans essential gene network based on phenotypic profiling of a complex tissue,” *Cell*, vol. 145, no. 3, pp. 470–482, 2011.

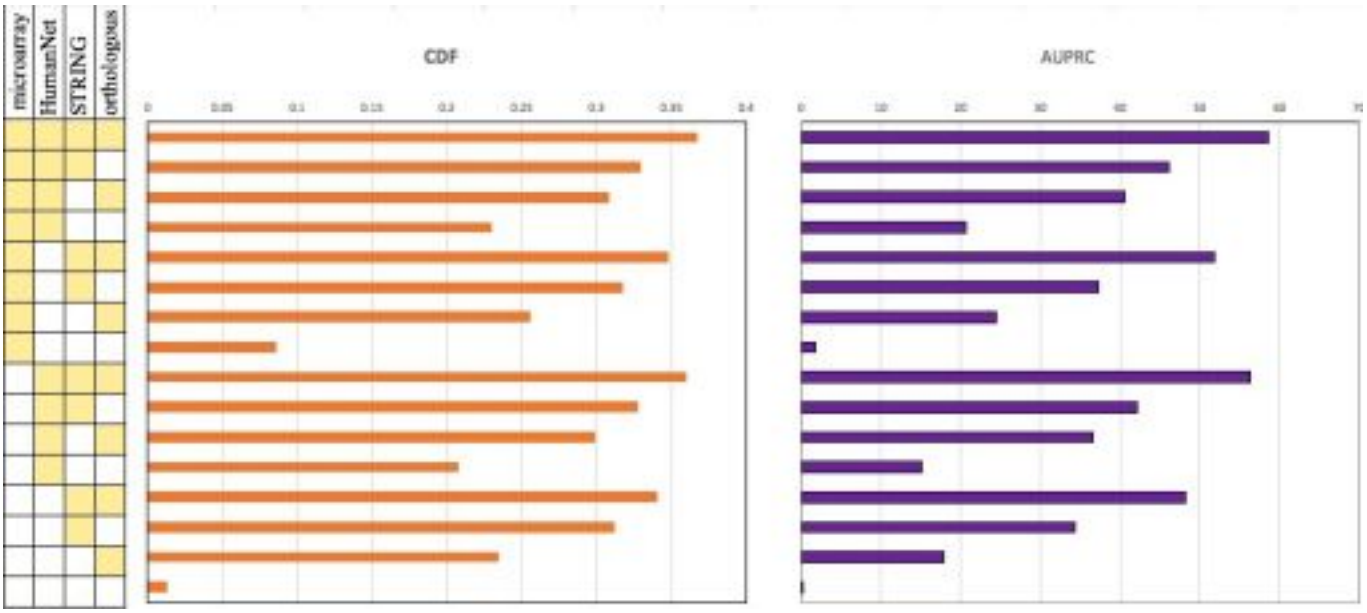


Fig. 5: Experimental results for evaluation of the importance of gene features. The grids in yellow indicate that the corresponding features are selected.

tions,” *Nucleic Acids Research*, vol. 37, tidak. Database issue, p. D555, 2009.

[28] JT Eppig, JA Blake, CJ Bult, JA Kadin, and JE Richardson, “The mouse genome database (mgd): new features facilitating a model system,” *Nucleic Acids Research*, vol. 35, tidak. Database issue, pp. 630–7, 2007.

[29] SS Dwight, MA Harris, K. Dolinski, CA Ball, G. Binkley, KR Christie, DG Fisk, L. Issel-Tarver, M. Schroeder, and G. Sherlock, “Saccharomyces genome database (sgd) provides secondary gene annotation using the gene ontology (go),” *Nucleic Acids Research*, vol. 30, tidak. 1, pp. 69–72, 2002.

[30] TL Saito, M. Ohtani, H. Sawai, F. Sano, A. Saka, D. Watanabe, M. Yukawa, Y. Ohya, and S. Morishita, “Scmd: Saccharomyces cerevisiae morphological database,” *Nucleic Acids Research*, vol. 32, tidak. 1, pp. 319–22, 2004.

[31] KL Mcgary, I. Lee, and EM Marcotte, “Broad network-based predictability of saccharomyces cerevisiae gene loss-of-function phenotypes.” *Genome Biology*, vol. 8, tidak. 12, hal. R258, 2007.

[32] ME Hillenmeyer, E. Fung, J. Wildenhain, SE Pierce, S. Hoon, W. Lee, M. Proctor, RP St Onge, M. Tyers, and D. Koller, “The chemical genomic portrait of yeast: uncovering a phenotype for all genes.” *Science*, vol. 320, tidak. 5874, pp. 362–365, 2008.

[33] RJ Nichols, S. Sen, YJ Choo, P. Beltrao, M. Zietek, R. Chaba, S. Lee, KM Kazmierczak, KJ Lee, and A. Wong, “Phenotypic landscape of a bacterial cell,” *Cell*, vol. 144, tidak. 1, pp. 143–156, 2011.

[34] J. Sprague, D. Clements, T. Conlin, P. Edwards, K. Frazer, K. Schaper, E. Segerdell, P. Song, B. Sprunger, and M. Westerfield, “The zebrafish information network (zfin): the zebrafish model organism database,” *Nucleic Acids Research*, vol. 34, tidak. 1, pp. 241– 243, 2003.

[35] GW Bell, TA Yatskievych, and PB Antin, “Geisha, a whole mount in situ hybridization gene expression screen in chicken embryos,” *Developmental Dynamics*, vol. 229, no. 3, pp. 677–687, 2010.

[36] D. Szklarczyk, JH Morris, H. Cook, M. Kuhn, S. Wyder, M. Si monovic, A. Santos, NT Doncheva, A. Roth, P. Bork *et al.*, “The string database in 2017: quality-controlled protein–protein associ ation networks, made broadly accessible,” *Nucleic acids research*, p. gkw937, 2016.

[37] JK Huang, DE Carlin, MK Yu, W. Zhang, JF Kreisberg, P. Tamayo, and T. Ideker, “Systematic evaluation of molecular networks for discovery of disease genes,” *Cell systems*, vol. 6, tidak. 4, pp. 484–495, 2018.

[38] F. Mordelet and JP Vert, “Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples,” *Bmc Bioinformatics*, vol. 12, tidak. 1, pp. 389–389, 2011.

[39] S. Karni, H. Soreq, and R. Sharan, “A network-based method for predicting disease-causing genes,” *Journal of Computational Biology*, vol. 16, tidak. 2, pp. 181–189, 2009.

[40] M. Xie, T. Hwang, and R. Kuang, *Prioritizing Disease Genes by Bi Random Walk*. Springer Berlin Heidelberg, 2015.

[41] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, “Deep canonical correlation analysis,” in *International Conference on Machine Learn ing*, 2013, pp. 1247–1255.