Penggunaan Fitur Kimiafisik dan Posisi Atom untuk Prediksi Struktur Sekunder Protein

Toto Haryanto[1], Budiman Surya Ardi[2]
[1,2]Departemen Ilmu Komputer Fakultas MIPA Institut Pertanian Bogor (IPB)
email: totoharyanto@ipb.ac.id, budimansurya.a@gmail.com

***Abstrak*—Prediksi struktur sekunder protein adalah salah satu masalah yang sudah lama dibahas dalam bidang bioinformatika. Berbagai metode telah diterapkan namun masalah akurasi belum mencapai hasil yang maksimal. Penelitian ini dilakukan untuk membangun suatu model prediksi struktur sekunder protein dengan menggunakan decision tree dengan fitur kimiafisik dan posisi atom. Penentuan setiap kelas dalam proses klasifikasi struktur sekunder protein dalam penelitian ini berdasarkan pembelajaran terhadap pola masukan dalam proses pelatihan. Data diperoleh dari Protein Data Bank (PDB). Sementara informasi struktur sekunder protein diperoleh melalui alamat DSSP. Sejumlah 75809 *alpha-helix* (H), 41785 untuk *bheta-sheet* (E), dan 80346 untuk *coil* (C) digunakan sebagai data set pada penelitian ini. Pola masukan diperoleh melalui proses *sliding window* dari sekuen asam amino dengan ekstraksi fitur kimiafisik dan posisi atom. Model prediksi dengan *cross validation* tanpa fitur posisi atom menghasilkan nilai akurasi 90.49%, sedangkan untuk pengujian dengan *unknown data* akurasinya menurun menjadi 51.29%. Akurasi menggunakan fitur posisi atom sebesar 90.17% dengan *cross validation* dan 50.83% jika diujikan pada *unknown data*.**

***Kata Kunci*— asam amino, *decision tree*, kimiafisik, prediksi struktur protein, posisi atom**

I. PENDAHULUAN

Protein mempunyai struktur yang sangat kompleks. Protein ini terbentuk dari urutan asam amino dengan karakteristik berbeda-beda. Secara hierarki, struktur protein dapat dikelompokkan menjadi 4 struktur utama yaitu struktur primer, struktur sekunder, struktur tersier dan struktur kuartener [1]. Struktur primer merupakan urutan asam amino yang dihasilkan dari ikatan peptida. Struktur sekunder adalah rangkaian asam amino yang membentuk struktur membelit, melingkar, dan melipat. Bentuk struktur ini dikelompokkan menjadi struktur alpha-helix (H), beetha-sheet (B), dan coil (C). Adapun struktur tersier merupakan gabungan dari berbagai struktur sekunder yang terjadi setelah proses pelipatan (*folding*).

Peranan protein ini terlihat jelas setelah rangkaian asam amino melakukan pelipatan dalam bentuk 3 dimensi (3D) sebagai struktur tersier. Namun struktur tersier (3D) tersebut ditentukan oleh struktur sebelumnya baik struktur primer maupun struktur sekundernya. Oleh karena itu penentuan struktur sekunder protein menjadi salah satu kajian yang banyak dilakukan di bidang bioinformatika.

Untuk mendapatkan sebuah struktur dari protein ditentukan secara eksperimen. Menurut Albert *et al* [2] struktur protein dapat ditentukan dengan eksperimen melalui penggunaan X-Ray Crystallography dan Nuclear Magnetic Resonance (NMR) spectroscopy. Keduanya mampu menghasilkan struktur protein sampai dengan bentuk 3 dimensinya. Dengan teknik ini sangat memungkinkan ditemukannya struktur protein baru. Namun hal ini tentu sangat sulit dan membutuhkan biaya yang tidak murah. Oleh karena itu, dengan perkembangan teknologi komputasi, untuk mendapatkan sebuah struktur protein dapat dilakukan dengan membuat model prediksi. Salah satu teknik komputasi yang dapat digunakan untuk memprediksi struktur sekunder protein adalah teknik klasifikasi *decision tree* dengan algoritme C4.5 yang merupakan sepuluh algoritma terbaik di bidang data mining [3] .

Penelitian yang terkait dengan prediksi struktur sekunder protein dilakukan oleh Lakizadeh [4] dengan menambahkan Contact Number (CN) sebagai variabel masukan dan menggunakan *sliding window* dengan lebar 13. Hasil penelitian ini adalah melihat pengaruh CN dalam meningkatkan akurasi dengan menggunakan jaringan saraf tiruan propagasi balik. Penelitian ini membuktikan bahwa dengan menambahkan CN dapat meningkatkan akurasi dalam memprediksi struktur sekunder protein secara signifikan seiring penambahan dari data subsetnya.

Penelitan lainnya dilakukan dengan menerapkan model Support Vector Machine (SVM) sebagai klasifikasi dan ekstraksi fitur kimiafisik. Model ini juga menggunakan s*liding window* dan teknik *filtering* [5]. Hasil dari model ini menghasilkan nilai akurasi yang lebih tinggi dengan menerapkan teknik *filtering* yaitu sebesar 79.52 %. Adapun model tanpa menggunakan teknik *filtering* hanya mampu menghasilkan akurasi sekitar 77.40 %. Prediksi struktur protein sekunder juga pernah dilakukan dengan menggunakan Hidden Markov Model (HMM) untuk kasus data yang tidak seimbang [6].

Kedua Penelitian sebelumnya memanfaatkan teknik klasifikasi dan ekstraksi ciri yang berbeda. Berdasarkan penelitian tersebut, penelitian ini diajukan untuk membuat prediksi struktur sekunder protein dengan menerapkan algoritme C4.5 dengan ekstraksi fitur kimiafisik dan posisi atom. Kemudian penelitian ini juga akan menentukan s*liding window* yang optimal agar didapat akurasi yang baik. Hasil dari klasifikasi akan membentuk sebuah model untuk memprediksi struktur sekunder protein.

II. METODE PENELITIAN

A. Struktur Protein

Protein (asal kata protos dalam bahasa Yunani yang memiliki arti "yang paling utama) merupakan salah satu dari biomakromolekul elemen penyusun utama makhluk hidup yang dibentuk dari asam amino (monomer) [1]. Protein dibentuk dari proses sintesis protein yang dilakukan melalui tahapan transkripsi dan translasi. Protein merupakan rangkaian biomolekul raksasa yang sangat penting peranannya untuk makhluk hidup. Untuk mendapatkan protein harus melalui proses translasi dari RNA. Proses untuk mendapatkan RNA melalui proses transkripsi dari DNA. Adapun proses terbentuknya protein dapat dilihat pada Gambar 1.

Adapun perubahan susunan basa nukleotida dari DNA menjadi RNA kemudian protein diperlihatkan pada Gambar 2.

Sekuens asam amino yang terbentuk dari proses translasi merupakan struktur primer. Sementara Struktur sekunder merupakan label dari setiap asam amino yang terbentuk, bisa

Gambar. 1. Proses Pembentukan Protein dari Rantai DNA melalui proses transkripsi menjadi RNA dan translasi menjadi protein

berupa alpha-helix(H), betha-sheet(B) atau coil (C). Ketiga struktur terebut dapat dibuat dengan menggunakan *tools* Cn3D

```
DNA     : TAC CGC GGC TAT TAC TGC CAG GAA GGA ACT
RNA     : AUG GCG CCG AUA AUG ACG GUC CUU CCU UGA
Protein : Met Ala Pro Ile Met Thr Val Leu Pro Stop
```

Gambar. 2. Ilustrasi perubahan basa nucleotide DNA menjadi RNA kemudian menjadi protein [6]

atau Rasmol. Gambar 3 menunjukkan potongan struktur sekunder yang dibuat dengan menggunakan perangkat lunak Rasmol versi 2.7.4.2 [7]

Adapun struktur tersier dari protein terbentuk apabila protein tersebut telah mengalami pelipatan seperti yang disajikan pada Gambar 4 [7].

(a) (b) (c)

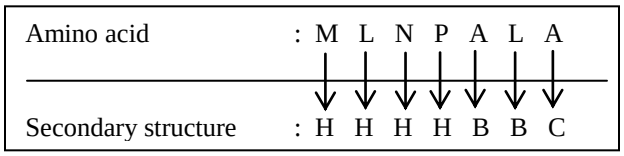
Gambar. 3. Visualisasi struktur sekunder protein (a) alpha-helix, (b) betha-sheet dan (c) coil.

Gambar. 4. Visualisasi salah satu struktur tersier protein menggunakan perangkat lunak Rasmol 2.7.4.2

B. Prediksi Struktur Sekonder Protein

Prediksi struktur sekunder protein menjadi riset yang menarik untuk dilakukan di bidang bioinformatika. Penelitian terkait yang dengan prediksi struktur protein banyak dilakukan karena berkaitan dengan fungsi dan peranan protein itu sendiri. Selain itu, pendekatan prediksi struktur protein berbasis komputasi ini merupakan pendekatan heuristic sehingga meskipun tidak akan mendapatkan struktur baru seperti layaknya menggunakan NMR atau kristalografi sinar-X namun diyakini membutuhkan waktu yang relatif lebih cepat.

Struktur sekunder protein merupakan label dari setiap asam amino yang membentuknya. Sebagai ilustrasi, Gambar 5 menunjukkan ilustrasi prediksi struktur protein sekunder.



Gambar. 5. Ilustrasi Prediksi Struktur Sekunder Protein

Dalam konteks klasifikasi, struktur sekunder H, B dan C merupakan kelas target sementara asam amino akan memberikan informasi sebagai penciri untuk memprediksi struktur sekunder tersebut.

C. Fitur Kimiafisik dan Posisi Atom

Penelitian ini akan menggunakan fitur kimiafisik dan posisi atom di dalam melakukan prediksi struktur sekunder protein. Untuk mendapatkan fitur tersebut dilakukan teknik yang dikenal dengan *sliding window.*

Fitur kimiafisik yang akan menjadi penciri dalam penelitian ini antara lain : *conformation parameter, net charge, hydrophobic* dan *side chain mass* [5]. Nilai *conformation parameter* menunjukkan peluang setiap residu asam amino terhadap struktur sekunder. Dengan kata lain untuk setiap asam amino akan dihitung peluang berada pada struktur alpha-helix (H), peluang pada struktur betha-sheet (B) dan peluang pada struktur coil (C). Nilai *Conformation parameter* setiap amino acid S_{ij} didefinisikan melalui persamaan (1).

$$S_{ij} = \frac{a_{ij}}{a_i} \text{ dengan i = 1,2,3 … 20 dan j= 1,2,3} \quad (1)$$

Variabel *i* merupakan asam amino sedangkan *j* menunjukkan struktur sekunder protein H, B dan C. Tabel 1 menunjukkan nilai *conformation parameter* asam amino pada ketiga struktur sekunder protein. Fitur *net charge* diperoleh

Tabel 1.
Nilai *conformation parameter*

Asam amino	Alpha-helix	Betha-sheet	Coil
A	0.53	0.16	0.30
R	0.46	0.17	0.37
N	0.30	0.13	0.57
D	0.32	0.12	0.57
C	0.32	0.30	0.38
E	0.52	0.15	0.33
Q	0.49	0.17	0.34
G	0.19	0.16	0.65
H	0.36	0.22	0.41
I	0.41	0.35	0.23
L	0.48	0.24	0.28
K	0.43	0.16	0.41
M	0.43	0.22	0.35
F	0.39	0.29	0.32
P	0.20	0.08	0.72
S	0.30	0.18	0.52
T	0.30	0.23	0.46
W	0.43	0.28	0.29
Y	0.38	0.30	0.32
V	0.32	0.42	0.26

berdasarkan tabel indeks amino acid (AA Index). Terdapat lima asam amino yang memiliki nilai net charges baik positif maupun negatif. *Hydrophobic* dan *side chain mass* digunakan sebagai fitur di dalam prediksi struktur sekunder protein karena terkait dengan proses *folding*. Tabel 2 menyajikan nilai dari fitur tersebut.

Posisi atom merupakan salah satu konten yang terdapat di dalam *file* dssp. Posisi atom menunjukkan koordinat tiga dimensi x,y dan z dari atom atom setiap asam amino. Dengan demikian, terdapat Sembilan fitur dasar yang akan menjadi vektor input untuk membuat model klasifikasi.

Tabel 2.
Fitur kimia fisik sebagai prediktor dalam prediksi struktur sekunder protein

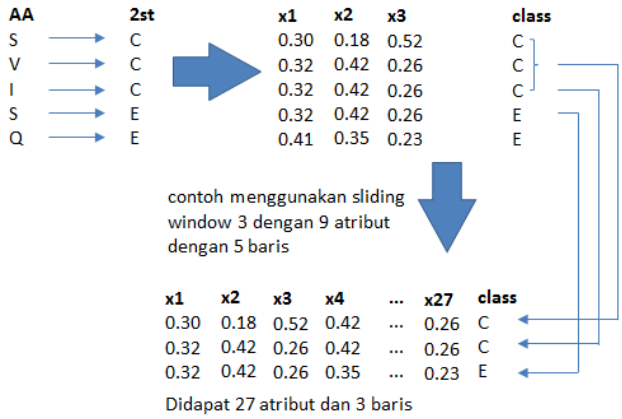
Asam amino	Net Charge	Hydrophobic	Side chain mass
A	0	1.8	15.0374
R	+1	-4.5	100.1431
N	0	-3.5	58.0597
D	-1	-3.5	59.0445
C	0	2.5	47.0947
E	-1	-3.5	73.0713
Q	0	-3.5	72.0865
G	0	-0.4	1.0079
H	+1	-3.2	81.0969
I	0	4.5	57.1151
L	0	3.8	57.1151
K	+1	-3.9	72.1297
M	0	1.9	75.1483
F	0	2.8	91.1323
P	0	-1.6	41.0725
S	0	-0.8	31.0341
T	0	-0.7	45.0609
W	0	-0.9	130.1689
Y	0	-1.3	107.1317
V	0	4.2	43.0883

D. Sliding window

Setelah Sembilan fitur tersebut diperoleh proses berikutnya adalah mempersiapkan data pelatihan. Teknik yang dilakukan pada tahap ini dikenal dengan *sliding window*. Pada penelitian ini, akan digunakan beberapa nilai lebar window (w). Nilai w yang digunakan berupa angka ganjil dengan posisi di tengah sebagai *point of interest* dan sekaligus sebagai label kelas dalam konteks kasus klasifikasi. Proses *sliding window* pada penelitian ini akan menjadi skenario dalam pembentukan model klasifikasi dan akan berdampak pada panjang vektor input model klasifikasi yang akan digunakan. Gambar 6 merupakan ilustrasi proses *sliding window* dengan lebar window (w) = 3.

Penelitian ini menggunakan beberapa nilai w mulai 7, 9, 11, 13, 15, 17 dan 19. Dengan demikian akan menggunakan beberapa variasi jumlah atribut untuk mendapatkan model klasifikasi yang paling baik.

E. Algoritma C.45



Gambar. 6. Ilustrasi proses sliding windw dengan lebar window (w) =3 sehingga menghasilkan jumlah fitur sebanyak 9 x 3 = 27 fitur

Algoritma ini digunakan untuk proses prediksi struktur prediksi sekunder protein. Algoritma ini digunakan untuk membangkitkan decision tree berdasarkan data latih yang disediakan. Algoritme ini merupakan pengembangan dari Algoritme ID3. Bahkan dalam suatu survey paper tahun 2008, C.45 termasuk 10 Top algorithm in data mining [8]. Algoritme ini pertama kali diperkenalkan tahun 1993 oleh Qunlan [9]. Algoritme C.45 dilakukan pada penelitian ini untuk melakukan klasifikasi struktur sekunder protein. Beberapa pengembangan yang dilakukan pada C4.5 adalah bisa mengatasi *missing value* bisa mengatasi *continuous* data dan prunning. Algoritme C4.5 menggunakan information gain untuk menetukan node akarnya. Secara umum algoritme C4.5 untuk membangun pohon keputusan adalah sebagai berikut:

1. Pilih atribut sebagai akar
2. Buat cabang untuk masing-masing nilai atribut
3. Bagi kasus dalam cabang
4. Ulangi proses untuk masing-masing cabang sampai semua kasus pada cabang memiliki kelas yang sama

Untuk memilih atribut masing-masing sebagai akar, didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung *gain* digunakan persamaan (2).

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} Entropy(S_i) \quad (2)$$

dengan
S : Himpunan kasus dari nilai A
A : Atribut
n : Jumlah partisi atribut A
$|S_i|$: Jumlah kasus pada partisi ke-i
$|S|$: Jumlah kasus dalam S

Sementara untuk menghitung nilai entropy dapat dilihat pada persamaan (3).

$$Entropy(S) = -\sum_{i=1}^{n} pi.\log 2.pi \quad (3)$$

dengan
S: Himpunan kasus
p: Proporsi dari S_i terhadap S

Formulasi untuk ratio gain dapat dilihat pada (4).

$$Gain\ Ratio\ (A) = \frac{Gain(A)}{Splitinfo(A)} \quad (4)$$

Dengan nilai Splitinfo didapat dari persamaan (5).

$$Splitinfo\ (A) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} log2 \frac{|S_i|}{|S|} \quad (5)$$

Prediksi struktur sekunder dengan Algoritme C.45 pada penelitian ini diimplementasikan menggunakan perangkat lunak Weka [10].

F. Evaluasi Hasil Prediksi

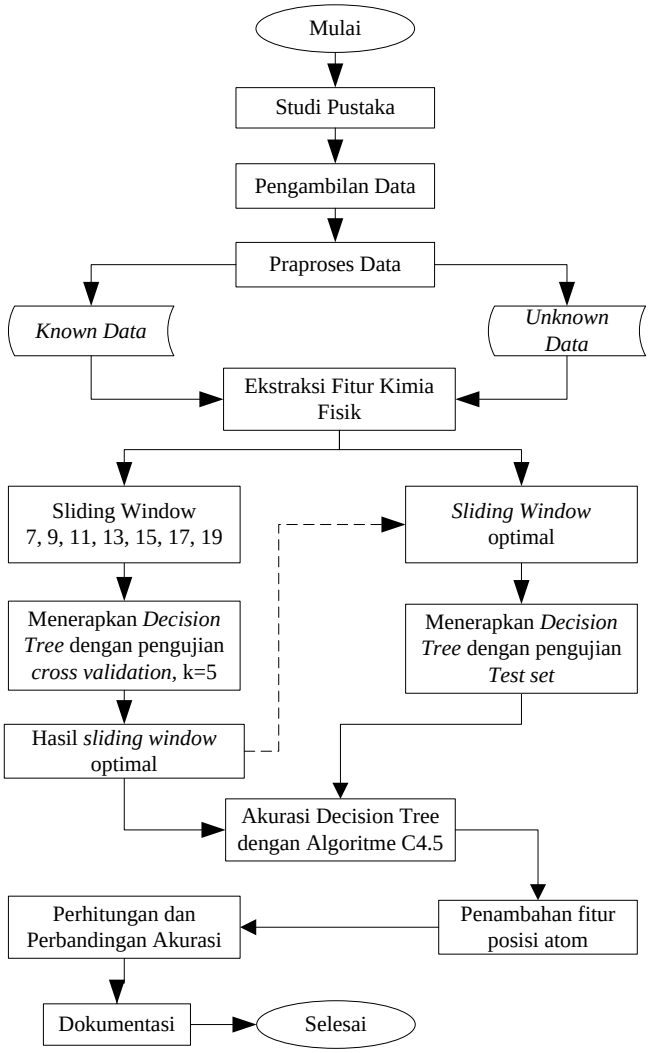
Pada tahap ini dilakukan pengujian terhadap model yang telah didapatkan dari proses pelatihan. Pengujian dilakukan dengan cara menghitung nilai Q3 *score*. Adapun perumusan dari perhitungan dari pengujian tersebut dapat dirumuskan pada persamaan (6)

$$Q_3 = \frac{N_H + N_B + N_C}{N_{Total}} \quad (6)$$

Dengan Q3 adalah akurasi rata-rata dari seluruh kelas, N_H adalah akurasi untuk alpha-helix, N_E adalah akurasi untuk bheta-sheet, dan N_C adalah akurasi untuk coil, dan N_{total} adalah jumlah kelas pada data uji.

G. Proses Prediksi Struktur Sekunder Protein

Secara umum proses prediksi struktur sekunder protein disajikan pada gambar 7 di bawah ini.

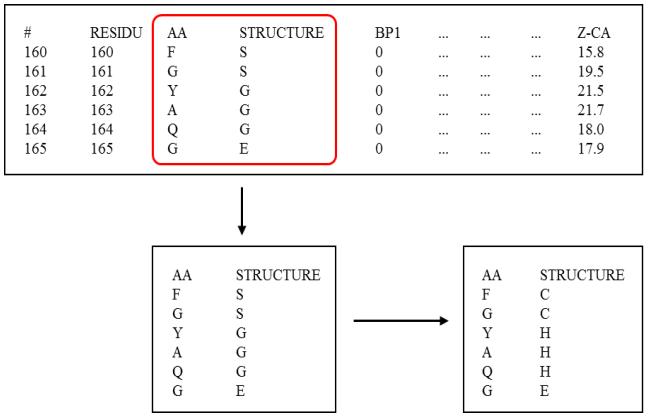


Gambar 7. Proses prediksi struktur sekunder protein

III. HASIL DAN PEMBAHASAN

A. Praproses Data

Praproses dilakukan untuk mempersiapkan data sebelum diolah ke dalam mesin pengklasifikasi. Di sisi lain, data ang diperoleh dari format *file* yang berekstensi DSSP masih dibutuhkan proses *parsing* sehingga informasi yang relevan saja yang akan diambil. Hasil praproses data digambarkan seperti pada Gambar 8.



Gambar 8. Ilustrasi tahap praproses untuk mendapatkan kelas label hasil *file* berekstenti .dssp

Pada setiap *file*, diambil data sekuen asam amino yang terdapat pada kolom {AA} serta pasangan data struktur sekundernya pada kolom {STRUCTURE}. Selain itu dilakukan reduksi data pada struktur sekunder yang memiliki 8 struktur sebelumnya menjadi 3 struktur sekunder [5]. Adapun hasil dari proses reduksi ini yaitu {I,H,G} menjadi alpha-helix(H), {E,B} menjadi betha-sheet (B) dan sisanya menjadi coil (C).

Data yang digunakan dalam penelitian merupakan data DSSP kategori enzyme comision yaitu hydrolases, isomerases, ligases, lyases, oxidoreductases, dan transferases. Data ini terdiri data 300 *file* known data dan 180 *file* unknown data. Dari known data yang digunakan terdapat 197940 record data struktur sekunder protein yang terdiri dari 75809 alpha-helix (H), 41785 untuk bheta-sheet (E), dan 80346 untuk coil (C). Kemudian pada cross validation dengan k adalah 5, total known data akan dibagi menjadi 5 bagian yang masing-masing bagiannya merupakan data uji dan sisanya data latih. Adapun untuk unknown data terdapat data struktur sekunder protein sebanyak 98402 record yang terdiri dari 41830 alpha-helix (H), 17478 untuk bheta-sheet (E), dan 39094 untuk coil (C). Persentase dari sebaran *known data* dan *unknown data* yang digunakan dapat dilihat pada Tabel 3.

Tabel 3.
Distribusi data setiap kelas alpha-helix(H), betha-sheet(B) dan coil (C)

	Kelas (%)		
	H	E	C
known data	40%	20%	40%
unknown data	42%	18%	40%

B. Hasil Ekstraksi Ciri

Penelitian ini menggunakan fitur kimiafisik dan posisi atom. Secara keseluruhan jumlah fitur kimiafisik adalah sebanyak enam fitur dan posisi atom sebanyak tiga fitur. Dengan demikian akan ada Sembilan fitur dasar sebagai prediktor untuk prediksi struktur sekunder protein ini sebalum dilakukan proses *sliding window*. Tabel 4 mengilustrasikan fitur yang digunakan.

Tabel 4.
Fitur kimiafisik dan posisi atom yang digunakan

X-H	X-E	XC	NC	HP	SCM	X-CA	Y-CA	Z-CA
0.30	0.18	0.52	-0.80	0	0.02	12.80	-14.40	2.60
0.32	0.42	0.26	4.20	0	0.03	9.60	-14.90	4.60
0.32	0.42	0.26	4.20	0	0.03	8.60	-15.80	8.30
0.32	0.42	0.26	4.20	0	0.03	9.30	-19.50	8.90
...	...	...	...	...	...	...	...	...
0.41	0.35	0.24	4.5	0	0.045	9.6	-20.8	12.5

Fitur kimia fisik — Fitur posisi atom

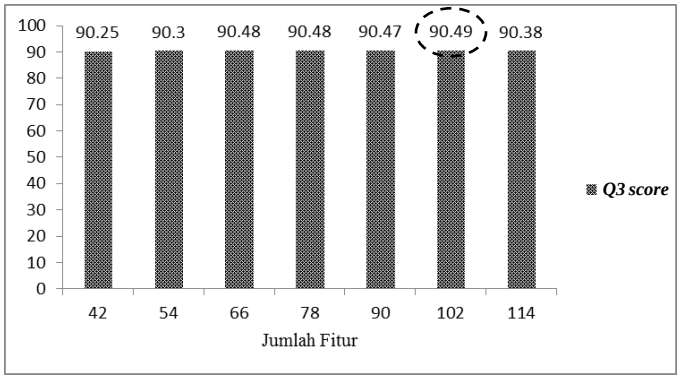
Setelah melakukan proses *sliding window*, maka akan didapatkan jumlah fitur yang bergantung pada lebar jendela (w) yang digunakan. Penelitian ini menggunakan jumlah fitur mulai dari 42, 54, 66, 78, 90, 102 dan 114 fitur. Adapun fitur posisi atom ditambahkan setelah diperoleh *sliding window* yang paling baik.

C. Pembentukan Model Klasifikasi dengan Decision Tree

Algoritme C.45 diterapkan untuk mendapatkan model klasifikasi untuk prediksi struktur sekunder protein. Perangkat lunak Weka digunakan untuk membuat model klasifikasi. Fitur yang masuk sebagai input pada proses pembentukan model bervariasi bergantung pada nilai (w) proses *sliding window*.

Hasil *sliding window* pada penelitian ini akan menghasilkan sekumpulan data untuk kemudian dijadikan sebagai data latih dan data uji. Adanya skenario beberapa *sliding window* ini digunakan untuk menghasilkan model paling baik yang ditandai dengan tingginya nilai Q3 *score* yang dihasilkan.

Pembentukan model pertama kali dilakukan dengan menggunakan data uji melalaui mekanisme *fold cross validation*. Penelitian ini menggunakan 5 *fold cross validation*. Setiap hasil prediksi kemudian dihitung nilai Q3 *score* yang didapatkan. Gambar 9 menunjukkan hasil prediksi menggunakan Algoritme C.45 dengan *5-fold cross validation*.



Gambar 9 Grafik perbandingan penggunaan jumlah fitur terhadap nilai Q3 *score* untuk prediksi struktur sekunder dengan skema *fold cross validation*

Hasil pengujian ini memperlihatkan bahwa Q3 *score* dari model yang terbaik dihasilkan dari penggunaan *sliding window* dengan ukuran 17 yang akan menghasilkan atribut

sebanyak 102 dan nilai *Q3* sebesar 90.49%. Adapun model yang menghasilkan akurasi Q3 terendah didapatkan dari penggunaan *sliding window* dengan ukuran 7 yaitu sebesar 90.25% dengan 42 atribut. Aturan dari model Decision Tree dengan *sliding window* 17 ini kemudian diuji dengan menggunakan *unknown data*. Hasil Q3 *score* dengan *unknown data* yang diperoleh sebesar 51, 29%.

Berdasar informasi nilai Q3 *score* tersebut, nilai w = 17 akan digunakan sebagai *window* yang paling baik untuk kemudian ditambahkan dengan fitur posisi atom. Hasil prediksi dengan menambahkan posisi atom menghasilkan jumlah fitur sebanyak 9 x 17 = 153 fitur.

Hasil Q3 *score* dengan penambahan fitur posisi atom ternyata menunjukkan nilai yang tidak lebih baik dibandingkan tanpa posisi atom yaitu sebesar 90,17%. Nilai ini juga diperoleh melalui mekanise *5-fold cross validation*. Sementara, ketika diujikan dengan *unknown data* menghasilkan Q3 *score* sebesar 50,83%. Detail hasil klasifikasi dapat dilihat pada Tabel 5.

Tabel 5.
Perbandingan hasil prediksi struktur sekunder protein dengan dan tanpa penambahan fitur posisi atom

	tanpa fitur posisi atom (%)				dengan fitur posisi atom (%)			
	H	E	C	Q3	H	E	C	Q3
Known data	91.09	88.28	91.08	90.49	90.83	87.87	90.74	90.17
Unknown data	51.06	32.56	54.48	51.29	57.89	32.32	53.70	50.83

Dari tabel di atas terlihat bahwa hasil prediksi struktur sekunder protein dengan dan tanpa fitur posisi atom tidak terlalu berbeda secara signifikan. Justru penggunaan posisi atom cenderung lebih mengurangi hasil nilai Q3 *score* hasil prediksi. Hal yang menarik juga dari hasil penelitian tersebut adalah bahwa ketika model diterapkan pada *unknown data* atau data yang baru, ternyata belum mampu untuk bisa menghasilkan hasil yang baik bahkan hanya mendapatkan hasil Q3 *score* sekitar 50 persen. Hasil ini memperlihatkan bahwa bisa jadi data yang baru tersebut memang sangat berbeda komposisi sekuens asam amino yang dimilikinya sehingga terjadi *overfitting*.

IV. KESIMPULAN/RINGKASAN

Decision tree mampu menghasilkan model yang cukup baik dalam memprediksi struktur sekunder protein jika menggunakan metode cross validation. Namun tidak baik dalam pengujian menggunakan *unknown* data.Nilai yang akurasi optimal dihasilkan dari penggunaan *sliding window* dengan ukuran 17. Penambahan fitur posisi atom sebagai informasi tidak mampu meningkatkan akurasi prediksi struktur sekunder protein.Adapun untuk penelitian berikutnya dapat disarankan untuk menambahkan fitur posisi matrik lain seperti PSSM atau protein contact number (CN).

DAFTAR PUSTAKA

[1] Polanski A, Kimmel M.2007. Bioinformatics. Germany (DE): Springer Science.

[2] Albert B *et al*. 1998. Essential Cell Biologi. An Introduction to the Molecular Biology of the Cell. New York (US). Garland Publishing, Inc

[3] Wu X dan Kumar V. 2008. The Top Ten Algorithm in Data Mining. CRC Press Taylor and Francis Group Boca Raton, US.

[4] Lakizadeh A, Marashi S. 2009. Addition of Contact Number Information can Improve Protein Secondary Structure Prediction by Neural Networks. *EXCLI Journal*. 8:66-73

[5] Huang YF, Chen SY. 2013. Extraxing physicochemical features to predict protein secondary structure. The Scientific World Journal. 2013:1-8. doi : 10.1155/2013/347106

[6] Sari DP dan Haryanto T. 2014. Penerapan Algoritme Viterbi pada Hidden Markov Model (HMM) untuk Prediksi Struktur Sekunder Protein. Prosiding Seminar Nasional Teknologi Informasi XI, 2014. Univeristas Tarumanegara. A6: 34-40

[7] Haryanto T, Buono A, Nugroho AS. 2011. Pengembangan Hidden Semi Markov Model dengan Distribusi Durasi Empiris untuk Prediksi Struktur Sekunder Protein. Thesis. Institut Pertanian Bogor (IPB).

[8] X. Wu et al. 2008. Top 10 algorithms in data mining. Survey Paper. Knowl Inf Syst. DOI 10.1007/s10115-007-0114-2.

[9] Quinlan, J. R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

[10] G. Holmes, A. Donkin, and I.H. Witten. Weka: A machine learning workbench. In*Proc Second Australia and New Zealand Conference on Intelligent Information Systems*, Brisbane, Australia, 1994.

.