دانشگاه صنعتي اصفهان
دانشكده برق و كامپيوتر

سمينار دفاع از پايان نامه كارشناسي ارشد كامپيوتر – هوش مصنوعي

پيش بيني نقشه تماس پروتئين توسط روش ماشين گروهي

ارائه دهنده
نرگس حبیبی

استاد راهنما
دكتر محمد حسين سرائي

استاد مشاور
دكتر حسن كربكندي

ارديبهشت 1388

روند ارائه مطالب

- بيوانفورماتيک
- پروتئين
- مفاهيم محاسباتي پيش زمينه
- نقشه تماس پروتئين
- پيش بيني نقشه تماس
- روش پيشنهادي
- نتيجه گيري
- پيشنهادها
- دستاوردهاي پژوهشي تحقيق

روند ارائه مطالب

- **بیوانفورماتیک**
- پروتئین
- مفاهیم محاسباتي پیش زمینه
- نقشه تماس پروتئین
- پیش بیني نقشه تماس
- روش پیشنهادي
- نتیجه گیري
- پیشنهادها
- دستاوردهاي پژوهشي تحقیق

بيوانفورماتيک - معرفي

- تعريف
  - علمي بين‌رشته‌اي (زيست شناسي، شيمي، فيزيک، رياضيات و علوم کامپيوتر)
  - کار بر روي داده‌هاي وسيع، متنوع و پيچيده زيست‌شناسي
  - با هدف حل مسائل زيست‌شناسي در سطح مولکولي

- دو زمينه اصلي
  - ژنوميک: تجزيه و تحليل ژن هاي موجودات زنده
  - پروتئوميک: بررسي پروتئين هاي موجودات زنده

بیوانفورماتیک-اهداف

- اهداف
  - سازمان دهي داده ها
    - مثال: بانک داده پروتئین
  - توسعه ابزارهائي براي کمک به تحلیل داده ها
    - مثال: مقایسه یک پروتئین با دیگر پروتئین ها
  - کشف الگوها و تفسیر نتایج
    - مثال: پیش بیني ساختار پروتئین

بیوانفورماتیک-کاربردها

- کاربردها
  - در زیست شناسي
    - یافتن ژن ها در میان ژنوم
    - پیش گوئي ساختار و عملکرد محصولات ژني
    - یافتن ارتباط میان ژن ها و توالي هاي پروتئین
  - در علوم پزشکي
    - طراحي منطقي داروها
    - شناسائي علل ژنتیکي بیماري ها

روند ارائه مطالب

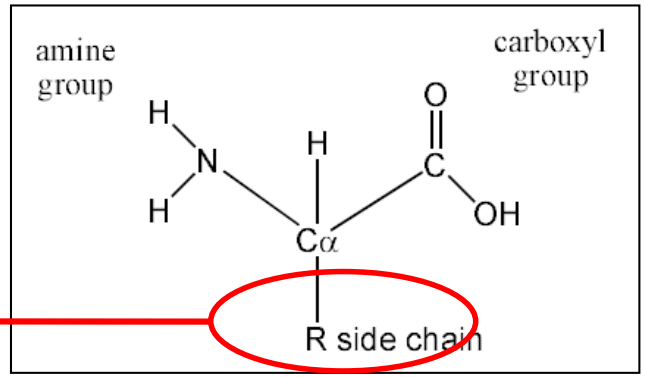
- بیوانفورماتیک
- **پروتئین**
- مفاهیم محاسباتي پیش زمینه
- نقشه تماس پروتئین
- پیش بیني نقشه تماس
- روش پیشنهادي
- نتیجه گیري
- پیشنهادها
- دستاوردهاي پژوهشي تحقیق

پروتئين

- تعريف: مولكولي مركب، متشكل از زنجيره اي از اسيدهاي آمينه
- ضرورت: مشاركت در هر پروسه درون سلولي
- نقش ها
  - آنزيمي: كاتاليز واكنش هاي مورد نياز موجود زنده
  - ساختاري
    - ديواره سلول
    - غشاي سلول
    - سيتوپلاسم
    - ...

پروتئين (ادامه)

- اسيدهاي آمينه
  - انواع: 20 نوع
  - ساختار



زنجيره
جانبي

ساختار عمومي يک اسيد آمينه

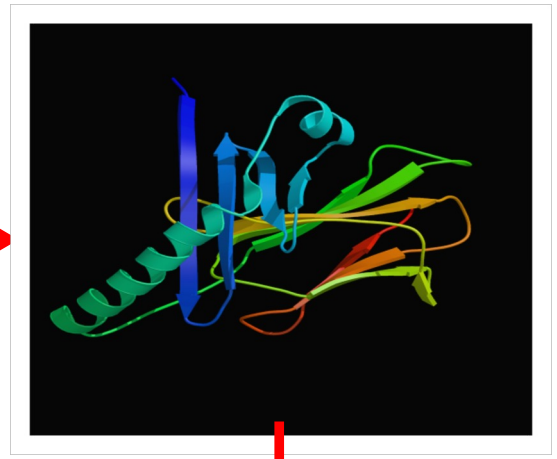
پروتئين (ادامه)

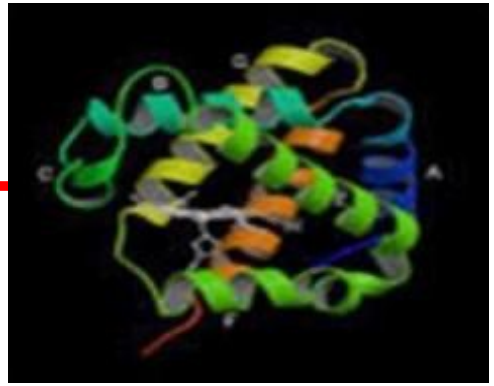
- سطوح مختلف ساختار پروتئين
  - ساختار اول: توالي اسيدهاي آمينه
  - ساختار دوم: مارپيچ هاي آلفا، صفحات بتا و ديگر ساختارهاي دوم
  - ساختار سوم: حاصل تا شدن زنجيره اسيدهاي آمينه، شامل چندين ساختار دوم
  - ساختار چهارم: اجتماع چند مولكول پروتئين

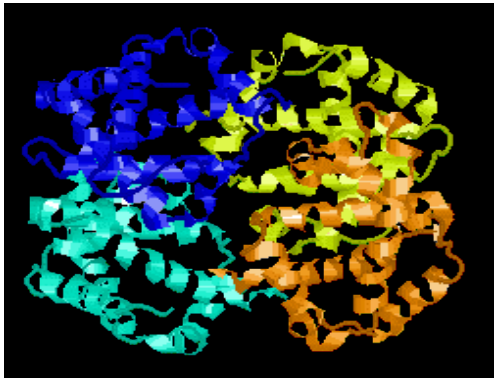
- نكته: وابستگي ساختارهاي دوم، سوم و چهارم پروتئين به ساختار اول آن

پروتئین (ادامه)

NH_2-Val-His-Leu-Thr-Pro-Glu-Glu-Lys-Ser-Ala-Ser-Ala-Val-Thr-Ala-Glu-His-…..







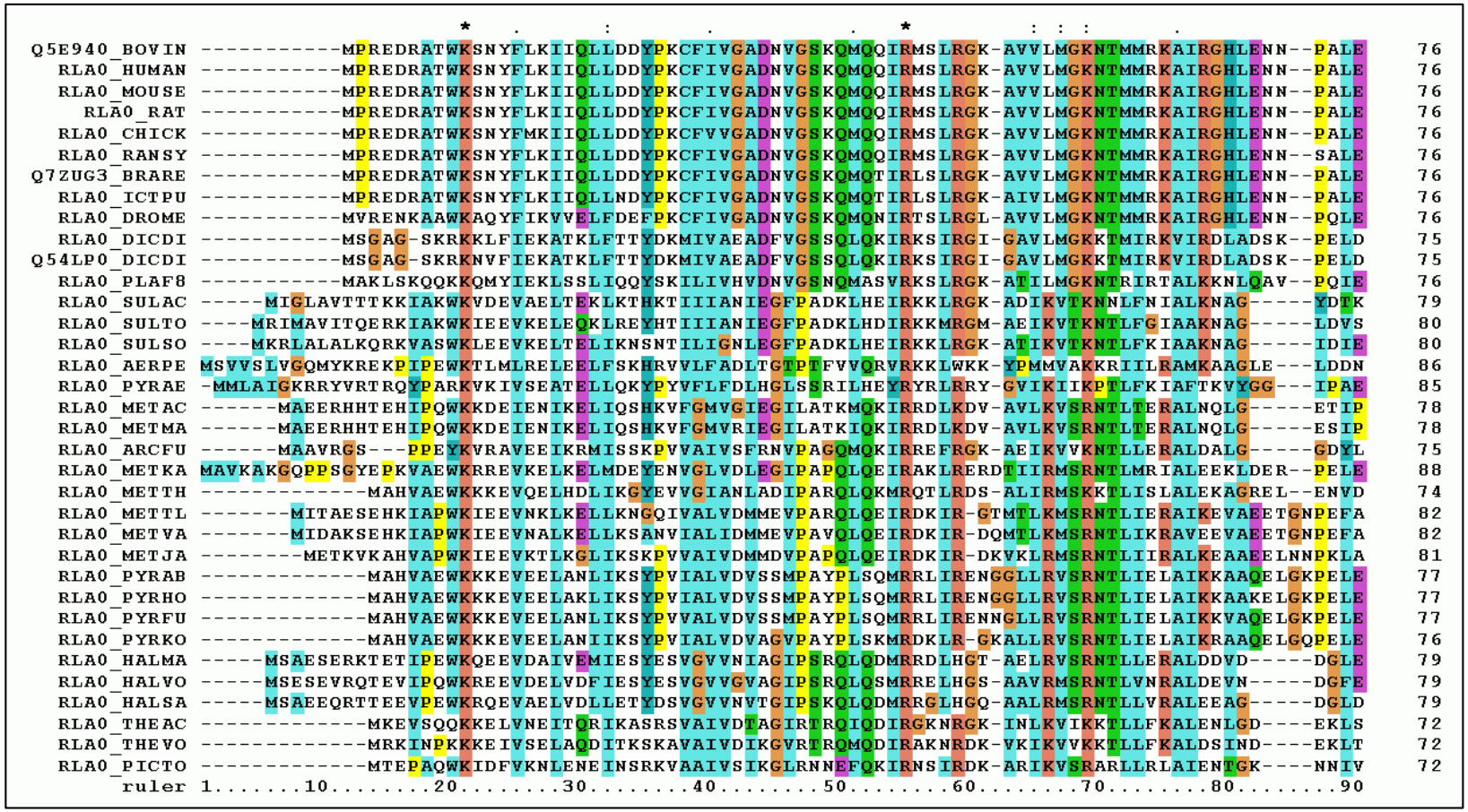
سطوح مختلف ساختار پروتئین

روند ارائه مطالب

- بیوانفورماتیک
- پروتئین
- **مفاهیم محاسباتي پیش زمینه**
- نقشه تماس پروتئین
- پیش بیني نقشه تماس
- روش پیشنهادي
- نتیجه گیري
- پیشنهادها
- دستاوردهاي پژوهشي تحقیق

مفاهيم محاسباتي پيش زمينه-MSA

- Multiple Sequence Alignment :MSA



بخشي از يک MSA براي چند توالي

مفاهيم محاسباتي پيش زمينه-اثر فراموشي خطرناک

- تعريف: از بين رفتن اطلاعات قبلي يادگرفته شده توسط شبکه عصبي، با يادگيري اطلاعات جديد

- علت: مناسب نبودن وزن هاي جديد شبکه براي تشخيص الگوهاي قبلي

- يک راه حل: تکرار

مفاهيم محاسباتي پيش زمينه-ماشين گروهي

- تعريف
  - گروهي از يادگيرها
  - آموزش يادگيرها بر روي يک مجموعه داده آموزشي
  - ترکيب خروجي‌هاي اعضا
- ايده: قاعده ((تقسيم کن و پيروز شو))
- مزيت ها:
  - بهتر بودن کارائي گروه از کارائي هر يک از اعضا
  - کاهش هزينه محاسباتي با تقسيم فضاي ورودي

مفاهيم محاسباتي پيش زمينه-ماشين گروهي (ادامه)

- علت استفاده: ضعف مدل هاي يادگيري ماشين در برخورد با مقادير زياد داده
  - رشد غيرخطي هزينه محاسباتي
  - اثر نامطلوب بر قابليت تعميم مدل
    - محدود بودن قابليت يک مدل
    - نياز به افزايش پارامترهاي مدل با افزايش تنوع الگوها
    - اثر نامطلوب افزايش پارامترها بر قابليت تعميم مدل و بر هزينه محاسباتي

مفاهيم محاسباتي پيش زمينه-ماشين گروهي (ادامه)

- انواع
  - ايستا
    - ميانگين گيري گروه
    - بوستينگ
  - پويا
    - تركيب يادگيرها
    - تركيب سلسله مراتبي يادگيرها

مفاهيم محاسباتي پيش زمينه-ماشين گروهي (ادامه)

- مثالي از روش ميانگين گيري گروه
  - گروهي از شبكه هاي عصبي با وزن هاي اوليه مختلف
  - آموزش شبكه ها با يک مجموعه داده
  - ميانگين گيري از پاسخ اعضا

- مزيت هاي روش ميانگين گيري گروه
  - افزايش كارائي
  - كاهش زمان آموزش
  - كاهش خطر بيش پوشش

روند ارائه مطالب

- بيوانفورماتيک
- پروتئين
- مفاهيم محاسباتي پيش زمينه
- **نقشه تماس پروتئين**
- پيش بيني نقشه تماس
- روش پيشنهادي
- نتيجه گيري
- پيشنهادها
- دستاوردهاي پژوهشي تحقيق

نقشه تماس پروتئین

- تعريف: ماتريسي دو بعدي، نشان دهنده وجود تماس يا عدم تماس بين هر دو اسيد آمينه در فضا

- مفهوم تماس: كمتر بودن فاصله بين دو اسيد آمينه در فضا از يک مقدار آستانه

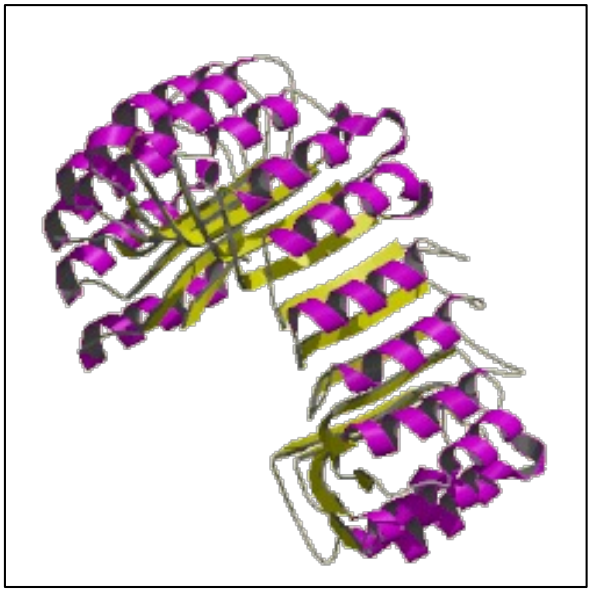
- بيان رياضي:

$$C = \begin{cases} 1 & D_{ij} < thr \\ 0 & otherwise \end{cases} \qquad i, j = 1, \ldots, N$$

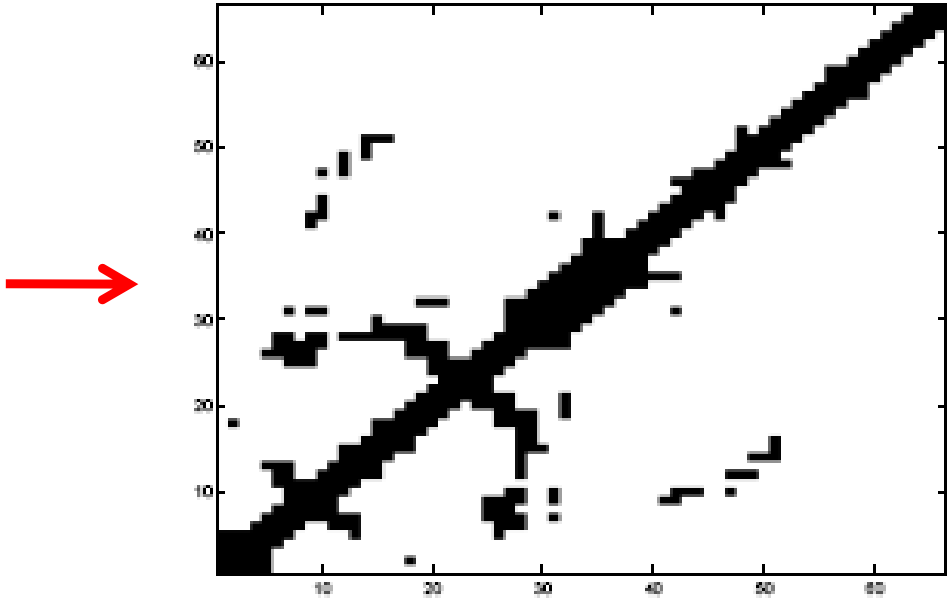
نقشه تماس پروتئين (ادامه)

- تعريف فاصله بين دو اسيد آمينه
  - مينيمم فاصله بين تمام اتم‌ها
  - فاصله بين اتم‌هاي C_α
  - فاصله بين اتم‌هاي C_β
  - ...
- ويژگي هاي نقشه تماس
  - محلي بودن بيشتر تماس ها
  - نسبت تعداد جفت هاي ((تماس دار)) به ((بدون تماس)): 1 به 20

نقشه تماس پروتئین (ادامه)



ساختار سوم یک پروتئین



نقشه تماس متناظر

روند ارائه مطالب

- بيوانفورماتيک
- پروتئين
- مفاهيم محاسباتي پيش زمينه
- نقشه تماس پروتئين
- **پيش بيني نقشه تماس**
- روش پيشنهادي
- نتيجه گيري
- پيشنهادها
- دستاوردهاي پژوهشي تحقيق

پيش بيني نقشه تماس

- هدف: پيش بيني نقشه تماس يک پروتئين براساس ساختار اول آن

- اهميت
  - وابسته بودن عمل پروتئين به ساختار سوم آن
  - هزينه بر بودن و زمان بر بودن روش هاي آزمايشگاهي تعيين ساختار سوم
  - نياز به روش هائي براي تعيين ساختار سوم، صرفا بر اساس ساختار اول
  - امكان پيش بيني ساختار سوم با داشتن نقشه تماس پيش بيني شده

پيش بيني نقشه تماس (ادامه)

- معيارهاي ارزيابي
  - دقت: درصد تماس هاي پيش بيني شده صحيح (TP)
  - پوشش: درصد تماس هاي واقعي پيش بيني شده
  - بهبود نسبت به پيش بيني كننده تصادفي
  - ...
- روش هاي ارائه شده
  - استفاده از مدل هاي يادگيري مختلف
  - استفاده از ويژگي هاي گوناگون
  - استفاده از معيارهاي متفاوت براي ارزيابي مدل ← مشكل بودن مقايسه روش ها

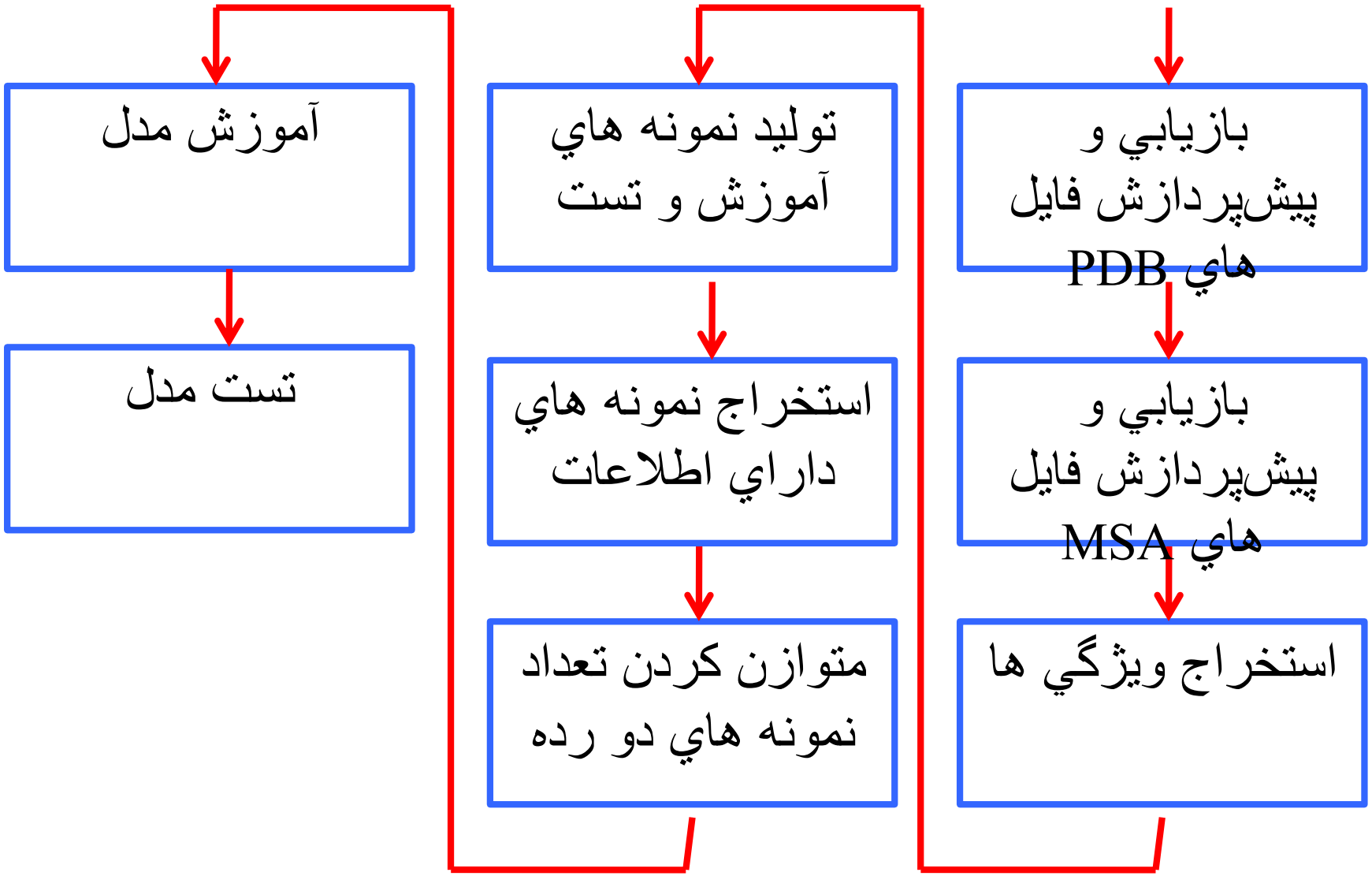
روند ارائه مطالب

- بيوانفورماتيک
- پروتئين
- مفاهيم محاسباتي پيش زمينه
- نقشه تماس پروتئين
- پيش بيني نقشه تماس
- **روش پيشنهادي**
- نتيجه گيري
- پيشنهادها
- دستاوردهاي پژوهشي تحقيق

روش پيشنهادي-معرفي كلي

- براساس روش ماشين گروهي
- ارائه يک ماشين گروهي جديد
- انتخاب شبکه هاي عصبي به عنوان اعضاي گروه
- ترکيب پاسخ اعضا به صورت محاسبه ميانگين پاسخ ها
- طراحي شبکه و ويژگي ها براساس روش Shackelford و Karplus
- در نظر گرفتن مساله پيش بيني نقشه تماس به عنوان يک مساله رده بندي

روش پيشنهادي-معرفي كلي (ادامه)



روش پيشنهادي-داده ها

- بازيابي فايل هاي داده از پايگاه داده PDB
- انتخاب فايل ها بر اساس ليست pdbselect-25%
- پيش پردازش فايل هاي PDB و استخراج مختصات اتم C_β هاي هر اسيد آمينه
- در نظر گرفتن 80% از فايل ها براي آموزش و 20% براي تست

روش پيشنهادي-MSA

- استخراج اكثر ويژگي ها از MSA پروتئين ها
- نحوه به دست آوردن: وب سرويس SAM-T08
- پيش پردازش فايل هاي به دست آمده

روش پيشنهادي-ويژگي هاي مورد استفاده

- ويژگي هاي توالي
  - طول توالي - فاصله توالي
- ويژگي هاي تک ستوني
  - توزيع اسيدهاي آمينه - آنتروپي - ساختار دوم پيش بيني شده
- ويژگي هاي جفت ستوني
  - تعداد جفت ها - آنتروپي مشترک – تمايل - MI E-value

روش پيشنهادي-آماده سازي داده ها

- شامل سه مرحله
  - توليد نمونه ها
  - استخراج نمونه هاي داراي اطلاعات
  - متوازن كردن نسبت نمونه هاي دو رده

روش پيشنهادي-آماده سازي داده ها (ادامه)

- توليد نمونه ها
  - در نظر گرفتن فاصله ميان اتم هاي C
  - آستانه تماس: 8 آنگستروم
  - آستانه تماس هاي محلي: 8
  - استفاده از پنجره اي به عرض 2 براي ويژگي هاي تک ستوني
  - استفاده از رتبه مقدار به جاي خود مقدار براي ويژگي هاي جفت ستوني
  - اندازه بردار ورودي شبکه هاي عصبي: 419

روش پيشنهادي-آماده سازي داده ها (ادامه)

- بردار ورودي
  - لگاريتم طول توالي
  - لگاريتم فاصله توالي
  - توزيع اسيد آمينه $i \pm 2$ (هر مكان 20 مقدار)
  - توزيع اسيد آمينه $j \pm 2$ (هر مكان 20 مقدار)
  - ساختار دوم پيش‌بيني شده $i \pm 2$ (هر مكان 21 مقدار)
  - ساختار دوم پيش‌بيني شده $j \pm 2$ (هر مكان 21 مقدار)
  - آنتروپي i
  - آنتروپي j
  - تعداد جفت‌هاي (i,j)
  - مقدار MI E-value (i,j)
  - لگاريتم رتبه MI E-value (i,j)
  - لگاريتم رتبه تمايل (i,j)
  - لگاريتم رتبه آنتروپي مشترک (i,j)
  - رده (i,j) (خروجي مطلوب)

روش پيشنهادي-آماده سازي داده ها (ادامه)

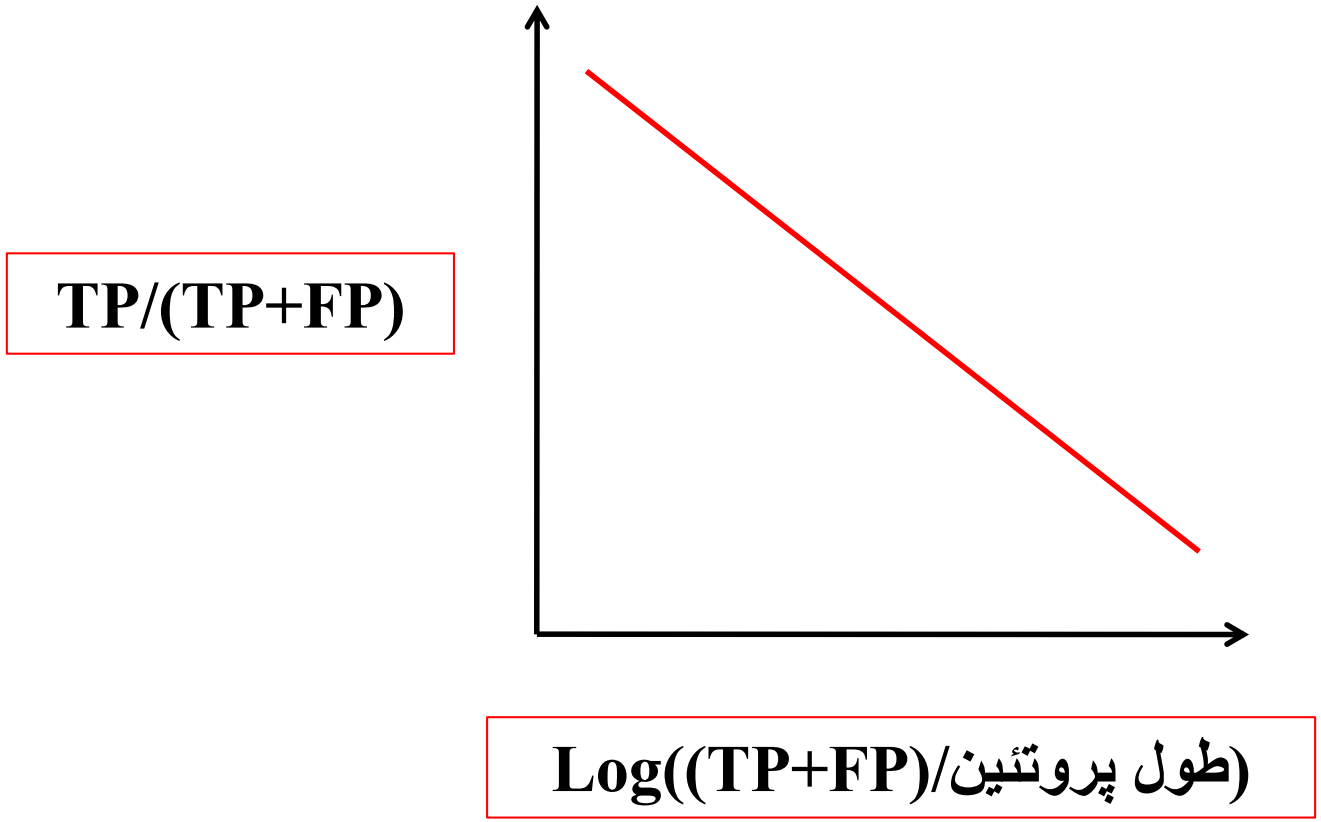
- استخراج نمونه هاي داراي اطلاعات
  - علت انجام
    - كاهش هزينه محاسباتي
    - كاهش مشكل عدم توازن تعداد نمونه هاي مثبت و منفي
  - توجيه: حذف نمونه هاي منفي بدون اثر بر روي مرز جداسازي دو رده
  - روش
    - مرتب سازي نمونه ها براساس چند ويژگي جفت ستوني
    - انتخاب 4L بالاترين نمونه ها از هر مجموعه (L طول پروتئين)
    - اجتماع نمونه هاي انتخاب شده از هر مجموعه

روش پيشنهادي-آماده سازي داده ها (ادامه)

- متوازن كردن نسبت تعداد نمونه هاي مثبت و منفي
  - علت: باياس شدن شبكه هاي عصبي به سمت رده منفي
  - نسبت متوازن سازي: 1 (مثبت) به 4 (منفي)
  - مشكل: ديده نشدن مقدار زيادي از الگوهاي رده منفي در حين آموزش
  - نكته: عدم متوازن نمودن داده هاي تست

روش پيشنهادي-نحوه ارزيابي كارائي

- معيار: نسبت تماس‌هاي پيش‌بيني شده صحيح به تعداد كل پيش‌بيني‌ها (TP/(TP+FP، در مقابل تعداد كل پيش‌بيني‌ها (TP+FP)



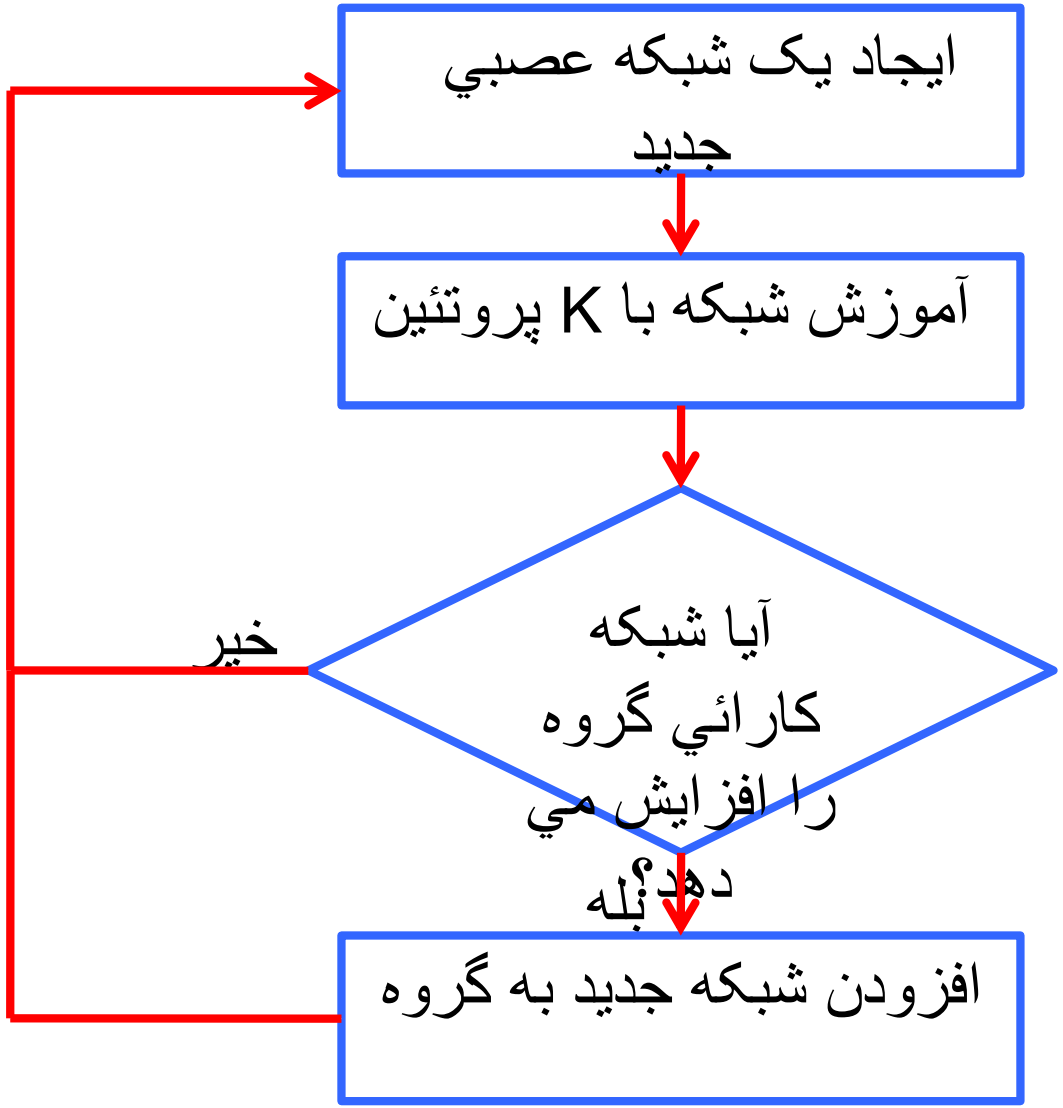
روش پيشنهادي-نحوه ارزيابي كارائي (ادامه)

- دو رويكرد موجود براي تعيين رده هر جفت اسيد آمينه
  - آستانه گذاري خروجي شبكه عصبي
  - ارزيابي كارائي برحسب آستانه هاي مختلف
- استفاده از رويكرد دوم
- نكته: نياز كاربردهاي مختلف به تعداد پيش بيني هاي صحيح متفاوت

روش پيشنهادي-آموزش مدل

- ايجاد يک ماشين گروهي ايستا در دو مرحله
  - مرحله اول: ايجاد گروه1
  - قرار دادن آخرين عضو گروه1 به عنوان اولين عضو گروه2
  - مرحله دوم: ايجاد گروه2 (گروه اصلي)

روش پيشنهادي-آموزش مدل (ادامه)



روش پیشنهادي-آموزش مدل (ادامه)

- ساختار شبکه هاي عصبي
  - feed-forward
  - یک لایه پنهان
  - تعداد نورون ها: 1-45-419
  - الگوریتم آموزش: بازگشت به عقب انعطاف پذیر
  - تابع کارائي: MSE
  - حداکثر تعداد epoch هاي آموزش: 500

روش پيشنهادي-آموزش مدل (ادامه)

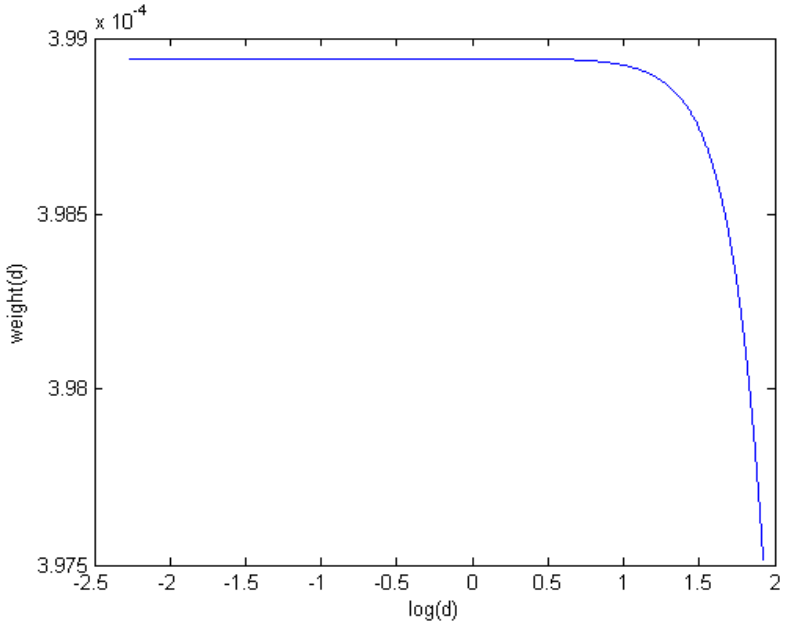
- آموزش هر شبکه با K پروتئين تصادفي (تقسيم فضاي ورودي)
  - چگونگي انتخاب مقدار K
- عدم آموزش يک شبکه با K پروتئين به طور سريال براي اجتناب از مشکل اثر فراموشي خطرناک
- استفاده از 90% داده هاي آموزشي براي آموزش و 10% براي جلوگيري از بيش پوشش

روش پيشنهادي-آموزش مدل (ادامه)

- نحوه تصميم گيري درباره اضافه شدن شبکه جديد به گروه
  - محاسبه دو نمودار کارائي
    - نمودار1: گروه جاري
    - نمودار2: گروه جاري + شبکه جديد
  - محاسبه نمودار تفاضل نمودار1و نمودار2
  - محاسبه مجموع وزن دار نمودار تفاضل
  - اضافه کردن شبکه جديد به گروه در صورت مثبت بودن مجموع

روش پيشنهادي-آموزش مدل (ادامه)

- تابع وزن پيشنهادي



تابع وزن پيشنهادي براي محاسبه
مجموع وزن دار نمودار تفاضل

روش پيشنهادي-آموزش مدل (ادامه)

- نكته: تقسيم مجموعه داده آموزشي به دو زيرمجموعه آموزش و تست
  - استفاده از زيرمجموعه آموزش براي آموزش هر شبكه جديد
  - استفاده از زيرمجموعه تست براي تصميم گيري در مورد اضافه شدن شبكه جديد به گروه
  - عدم استفاده از زيرمجموعه تست براي تست نهائي مدل

روش پيشنهادي-تست مدل

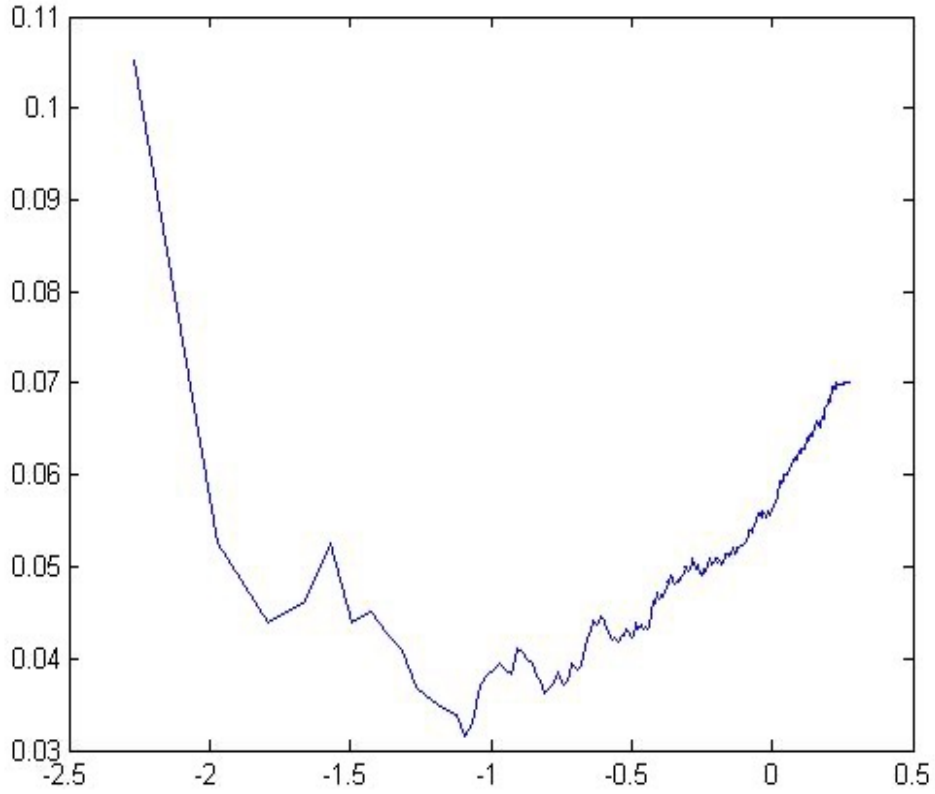
- كارائي مدل: ميانگين گيري از كارائي هاي مدل بر روي هر پروتئين تست

- پياده سازي سه روش پيش بيني
  - روش پايه: آموزش يک شبکه با تمام داده هاي آموزشي به طور سريال
  - روش ميانگين گيري گروه: آموزش 10 شبکه (با وزن هاي اوليه متفاوت) با تمام داده هاي آموزشي به طور سريال و ميانگين گيري از پاسخ شبکه ها
  - روش پيشنهادي: گروهي با 105 عضو

روش پيشنهادي-بررسي نتايج

- نتيجه روش پايه:

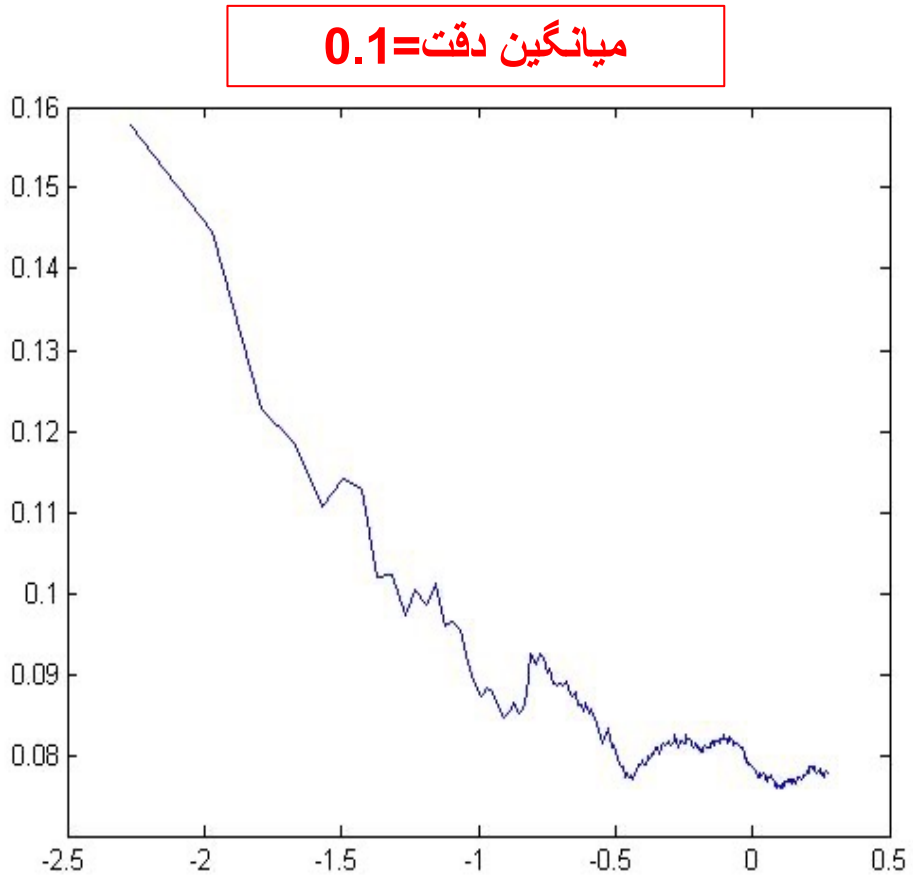
ميانگين دقت=0.05



نمودار كارائي روش پايه

روش پيشنهادي-بررسي نتايج (ادامه)

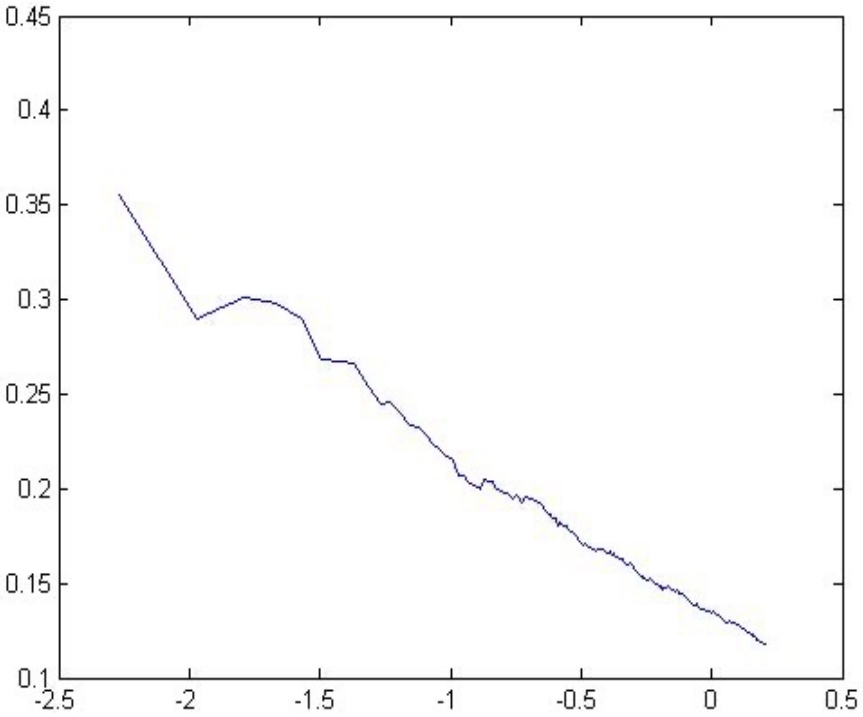
- نتيجه روش ميانگين گيري گروه



نمودار كارائي روش ميانگين‌گيري گروه

روش پيشنهادي-بررسي نتايج (ادامه)

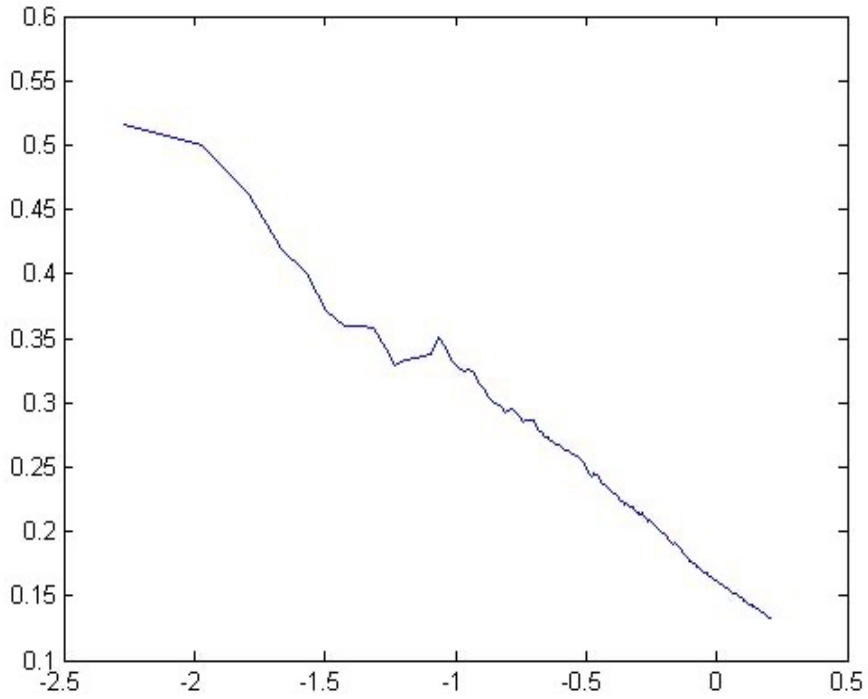
- نتايج روش پيشنهادي



نمودار كارائي گروه1 بر روي زيرمجموعه تست داده‌هاي آموزشي

روش پيشنهادي-بررسي نتايج (ادامه)

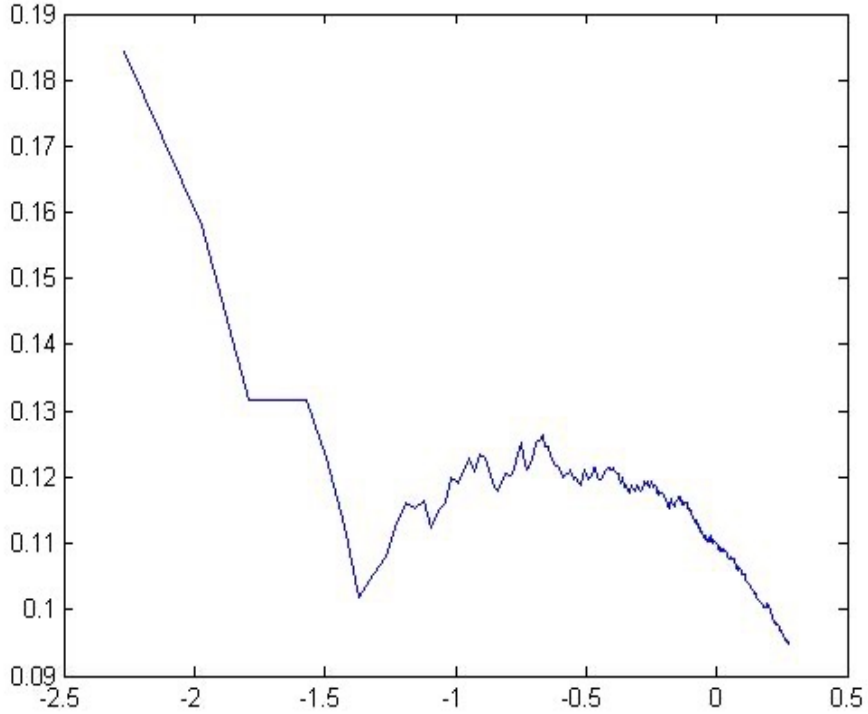
- نتايج روش پيشنهادي



نمودار كارائي گروه2 بر روي زيرمجموعه تست داده‌هاي آموزشي

روش پيشنهادي-بررسي نتايج (ادامه)

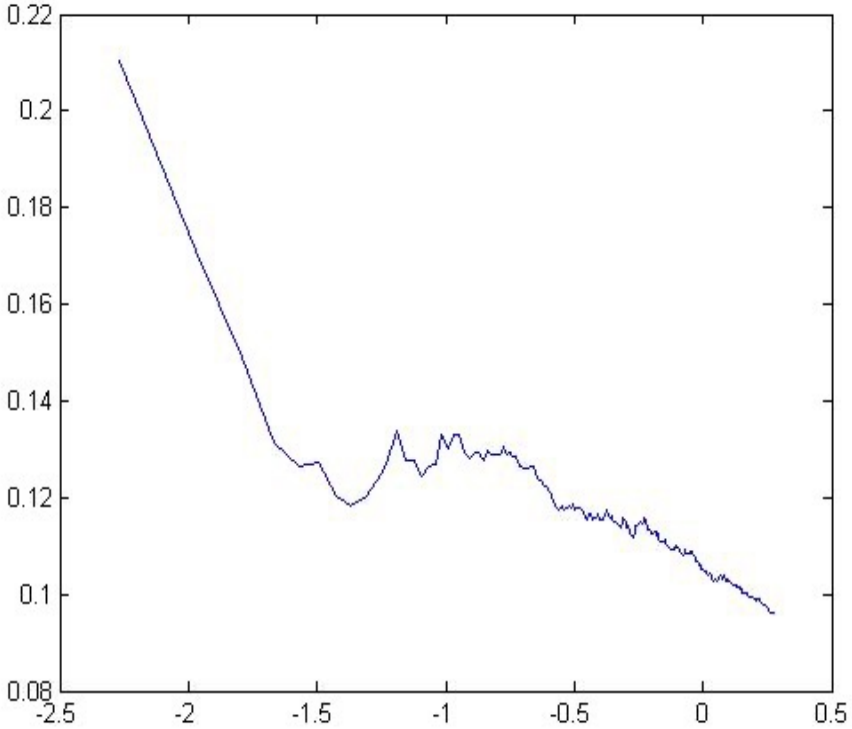
- نتايج روش پيشنهادي



نمودار كارائي گروه2 با 5 عضو

روش پيشنهادي-بررسي نتايج (ادامه)

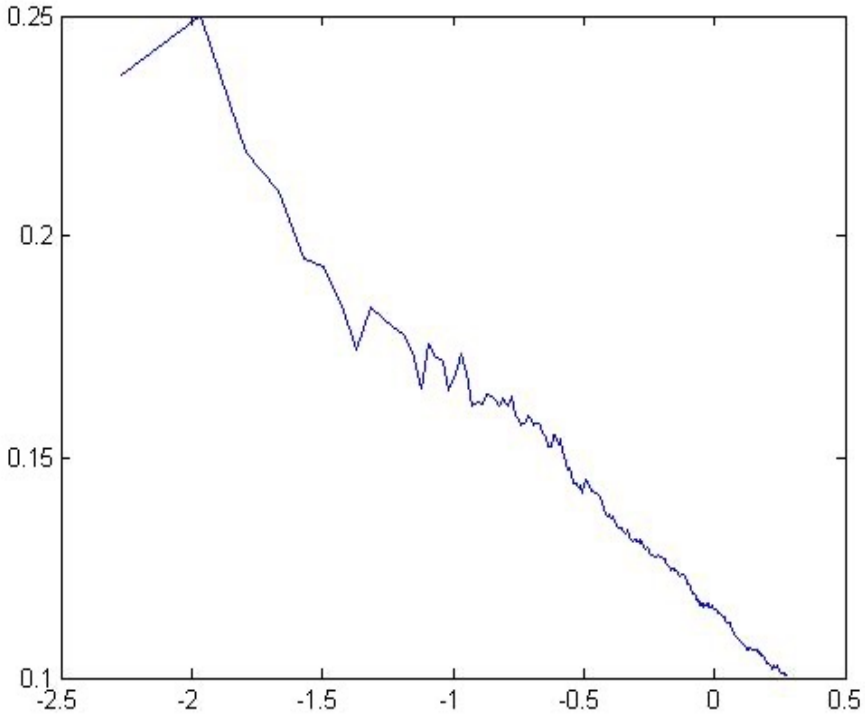
- نتايج روش پيشنهادي



نمودار كارائي گروه 2 با 20 عضو

روش پيشنهادي-بررسي نتايج (ادامه)

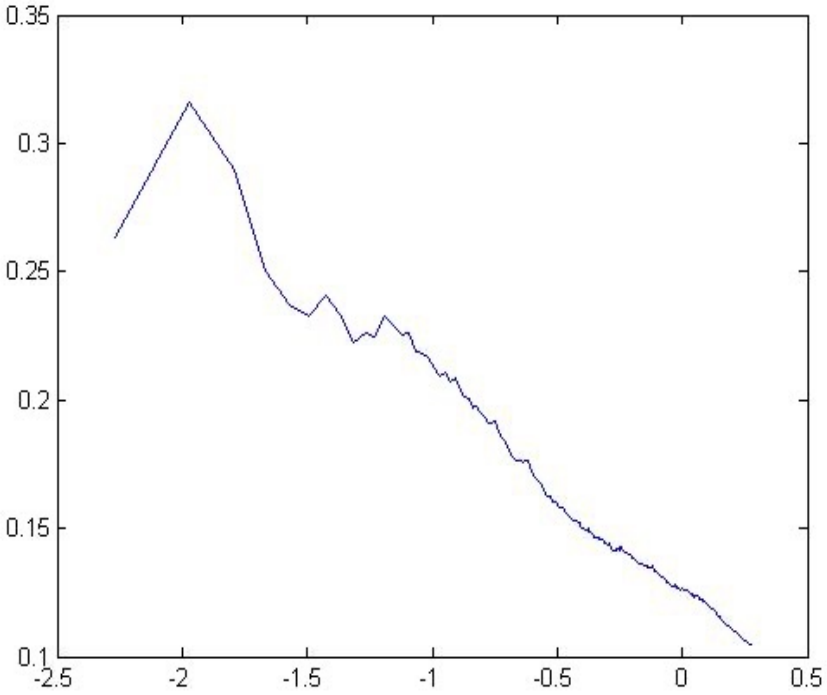
- نتايج روش پيشنهادي



نمودار كارائي گروه 2 با 50 عضو

روش پيشنهادي-بررسي نتايج (ادامه)

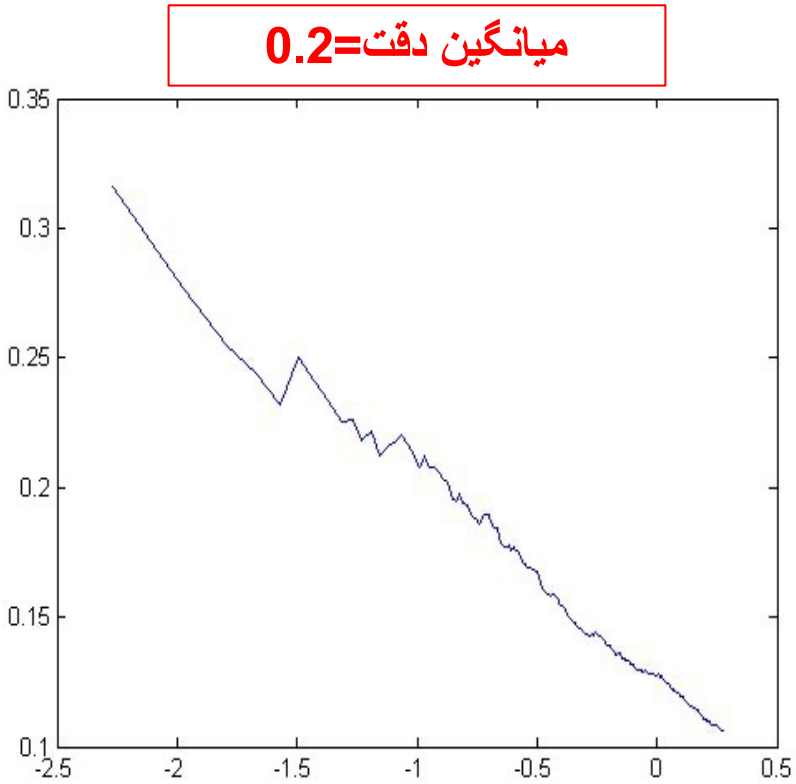
- نتايج روش پيشنهادي



نمودار كارائي گروه 2 با 90 عضو

روش پيشنهادي-بررسي نتايج (ادامه)

- نتايج روش پيشنهادي



نمودار كارائي مدل

روند ارائه مطالب

- بيوانفورماتيک
- پروتئين
- مفاهيم محاسباتي پيش زمينه
- نقشه تماس پروتئين
- پيش بيني نقشه تماس
- روش پيشنهادي
- **نتيجه گيري**
- پيشنهادها
- دستاوردهاي پژوهشي تحقيق

نتيجه گيري

- ارائه روشی نوين پيش بينی نقشه تماس بر مبناي تکنيک ماشين گروهي
- نشان دادن کارائي روش پيشنهادي با بررسي نتايج
- مزاياي روش پيشنهادي
  - کاهش خطا
  - کاهش واريانس
  - کاهش بيش پوشش
- معايب روش پيشنهادي
  - نياز به حافظه بيشتر
  - افزايش هزينه آموزش
  - پيچيده تر شدن تحليل سيستم

روند ارائه مطالب

- بيوانفورماتيک
- پروتئين
- مفاهيم محاسباتي پيش زمينه
- نقشه تماس پروتئين
- پيش بيني نقشه تماس
- روش پيشنهادي
- نتيجه گيري
- **پيشنهادها**
- دستاوردهاي پژوهشي تحقيق

پیشنهادها

- پیش پردازش متفاوت برخي فايل هاي PDB، به جاي حذف آن ها
- پیش پردازش فایل‌هاي MSA با دو آستانه شباهت متفاوت براي استخراج ویژگي‌هاي تک ستوني و جفت ستوني
- محاسبه ویژگي توزیع اسیدهاي آمینه به روش تنظیم کننده ترکیبي دریکله، به جاي محاسبه تقریبي احتمال رخداد آن‌ها
- استفاده از الفباها و روش‌هاي دیگر پیش بیني ساختار دوم
- استفاده از اندیس‌هاي AAindex3 به جاي ویژگي تمایل

پيشنهادها (ادامه)

- استفاده از انواع ديگر ماشين گروهي به جاي مدل پيشنهادي
- استفاده از مدل‌هاي يادگيري ديگر مانند ماشين بردار پشتيبان و مدل پنهان ماركوف، به جاي شبكه عصبي
- استفاده همزمان از چند نوع مدل يادگيري در ماشين گروهي
- تعريف معياري كاراتر براي تصميم گيري درباره افزودن شبكه جديد به گروه
- حذف تماس هاي غيرمحتمل از مجموعه تماس‌هاي پيش‌بيني شده، بر اساس قواعد بيوشيمي

روند ارائه مطالب

- بيوانفورماتيک
- پروتئين
- مفاهيم محاسباتي پيش زمينه
- نقشه تماس پروتئين
- پيش بيني نقشه تماس
- روش پيشنهادي
- نتيجه گيري
- پيشنهادها
- **دستاوردهاي پژوهشي تحقيق**

دستاوردهاي پژوهشي تحقيق

Protein Contact Map Prediction Based On an Ensemble Learning Method

Narjes Khatoon Habibi, Mohammad Hossein Saraee
Advanced Database Systems, Data Mining and Bioinformatics Research Laboratory, Department of Electrical and *Computer Engineering*
Isfahan University of Technology
Isfahan, Iran, 84156-83111
nhabibi@ec.iut.ac.ir, saraee@cc.iut.ac.ir

***Abstract*—Contact map is the simplified, 2D representation of protein spatial structure. Contact map prediction is an intermediate step to predict protein 3D structure. Ensemble learning-based model is a collection of learners that is more accurate than a single learner. In this paper we propose an ensemble learning method for contact map prediction. Results show that a considerable performance improvement is attainable using this approach instead of using a single model.**

I. INTRODUCTION

Proteins are the basic functional units of the cell which accomplish almost all the activities of life. The determination of the structure of proteins is an important step toward understanding the behavior of proteins. Experimental methods, such as x-ray crystallography and NMR techniques are not sufficient enough to allow for fast structural determination of the increasing number of newly discovered sequences. Therefor, employing computational techniques becomes vital.

Contact map is a simplified protein topology representation and it is an intermediate step to predict protein structure. it provides useful information about protein's structure. For example, secondary structure can easily be recognized from it. Alpha-helices appear as dense bands along the main diagonal since they involve contacts between one amino acid and its four successors, while Beta-sheets are dense bands parallel or anti-parallel to the main diagonal, etc [1].

Over the years, a variety of different approaches have been developed for contact map prediction, including statistical methods using correlated mutations [2,3], machine learning [1,4,5,6,7,8,9] and threading template-based voting [10].

The machine learning methods generate contact predictions by training the contact maps of known structures on a variety of sequence-based features including sequence profiles, secondary structure predictions, sequence conservation, correlated mutations and folding initiation sites (i. e. I-sites). Different thechniques of machine learning such as neural networks [4,5,10], support vector machines [6,7], evolutionlary algorithms [1], and association rule based classification [9] have been used in this field. However, the accuracy of contact map prediction is still far from perfect.

In This paper we try to get use of a more advanced technique named ensemble learning. It has been shown that by using this technique, the overall model would have a better precision and performance in comparison to the single learner methods.

II. DEFINITIONS

Background materials are presented in the following sub-sections.

A. Contact Map

Given a protein sequence P with N residues, P = {$R_1, R_2, \ldots, R_N$}, two residues R_i and R_j ($0< i, j \leq N$), are considered to be in contact if their spatial distance is less than a threshold *Ct*. The distance between two residues, has various definitions in the literature. It could be defined as the distance between $C_\alpha - C_\alpha$ atoms of the two residues [8], $C_\beta - C_\beta$ [10], and the minimal distance between atoms belonging to the side chain or the backbone of the residue pair [11]. In this work, the distances between $C_\beta - C_\beta$ are adopted. The distance function is [4]:

$$D_\beta = |\vec{C}_{\beta,i} - \vec{C}_{\beta,j}| = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (1)$$

Where $\vec{C}_{\beta,i} = (x_i, y_i, z_i)$ is the three-dimentional coordination of the i^{th} C_β atom (i.e. the C_β atom belonging to the i^{th} residue of the protein sequence).

We now could define a pairwise residue distance matrix $D = \{ D_{ij} \}$ for each protein, where each element C_{ij} of the contact map is set to 1 if the residue pair (R_i,R_j) is in contact, otherwise, C_{ij} is set to 0. In other words, the contact map of a protein with the length of N can be displayed by an $N \times N$ matrix, whose elements can be defined as:

دستاوردهاي پژوهشي تحقيق (ادامه)

Mining of Protein Primary Structure Data Using Committee Machines Approach to Predict Protein Contact Map

Narjes K. Habibi[1], Kaveh Mahdaviani[1,2] (IEEE Member), Mohammad H. Saraee[1]
[1]Advanced Database Systems, Data Mining and Bioinformatics Research Laboratory, Department of Electrical and Computer Engineering
Isfahan University of Technology (84156-83111)
[2]Isfahan Mathematics House
Isfahan, Iran
nhabibi@ec.iut.ac.ir, kaveh@ec.iut.ac.ir, saraee@cc.iut.ac.ir

Abstract—**Committee machines approach has shown to be useful in different applications. Protein primary structure data contain valuable information to extract. In this paper we mine these data and predict protein contact map based on committee machines. Contact map is the simplified, two dimensional representation of protein spatial structure. Contact map prediction is of great interest due to its application in fold recognition and predicting protein tertiary structure. The results show that the performance of the committee is considerably better than a single model.**

artificial intelligence; machine learning; committee machine; ensemble learning; neural network; bioinformatics; protein contact map; contact map prediction

I. Introduction

Proteins are the basic functional units of the cell which accomplish almost all the activities of life. The determination of the structure of proteins is an important step toward understanding the behavior of proteins. Experimental methods, such as x-ray crystallography and NMR techniques are not sufficient enough to allow for fast structural determination of the increasing number of newly discovered sequences. Therefor, employing computational techniques becomes vital.

Contact map is a simplified protein topology representation, and it is an intermediate step to predict protein structure. it provides useful information about protein's structure. For example, secondary structure can easily be recognized from it. Alpha-helices appear as dense bands along the main diagonal since they involve contacts between one amino acid and its four successors, while Beta-sheets are dense bands parallel or anti-parallel to the main diagonal, etc [1].

Over the years, a variety of different approaches have been developed for contact map prediction, including statistical methods using correlated mutations [2,3], machine learning [1,4,5,6,7,8,9] and threading template-based voting [10].

The machine learning methods generate contact predictions by training the contact maps of known structures on a variety of sequence-based features including sequence profiles, secondary structure predictions, seqeunce conservation, correlated mutations and folding initiation sites (i. e. I-sites). Different techniques of machine learning such as neural networks [4,5,10], support vector machines [6,7], evolutionlary algorithms [1], and association rule based classification [9] have been used in this field. However, the accuracy of contact map prediction is still far from perfect.

In This paper we try to get use of a more advanced technique named committee machines. It has been shown that by using this technique, the overall model would have a better precision and performance in comparison to the single learner methods.

The rest of this paper is organized as follows. Section II provides some background information on contact maps, catastrophic forgetting and committee machines. In section III, the proposed model and applied input features are described in more details. Section IV presents experimental results. Finally, Section V provides some concluding remarks and future works.

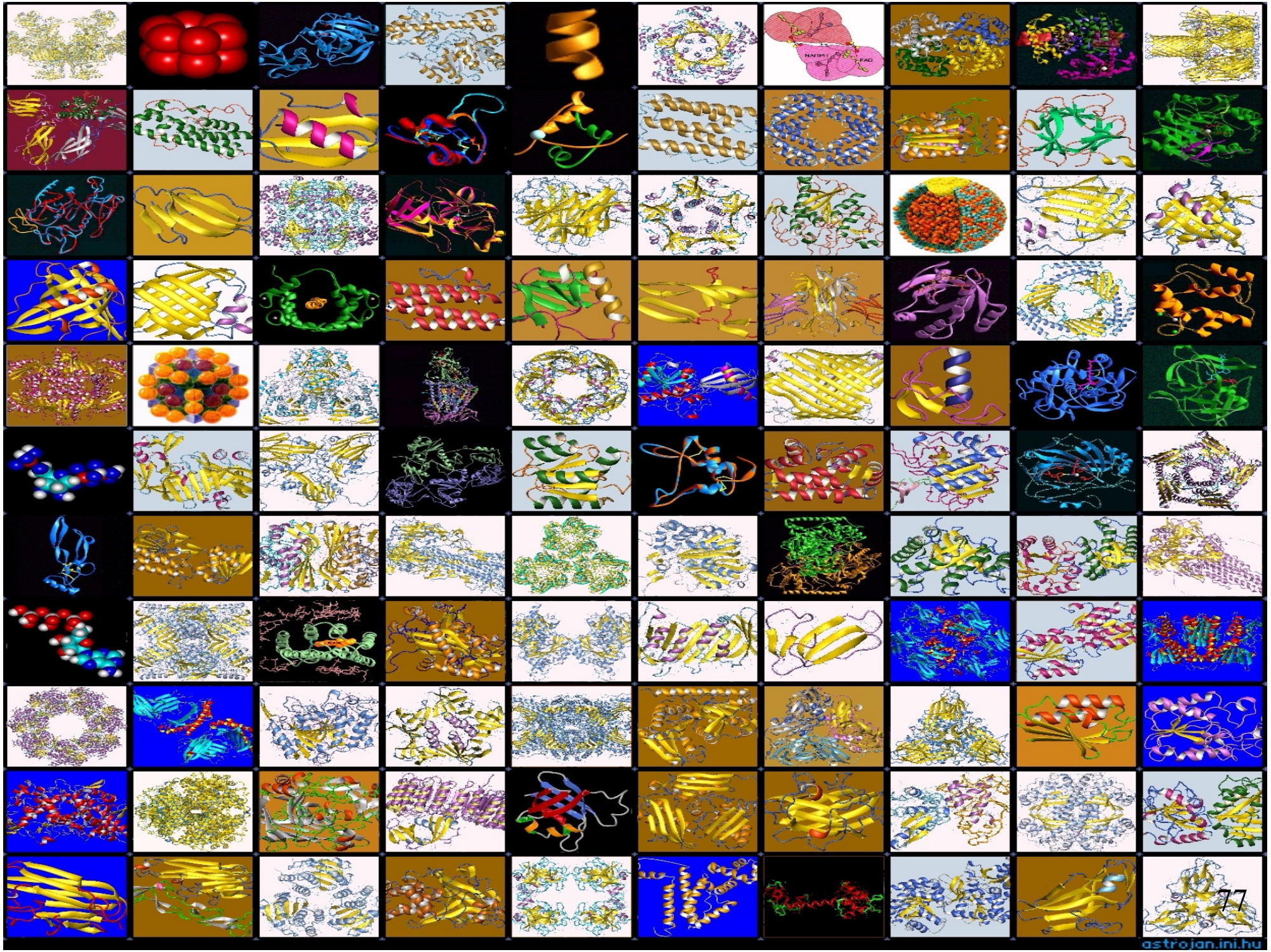
II. Background Materials

A. Contact Map

Given a protein sequence P with N residues, $P = \{R_1, R_2, \ldots, R_N\}$, two residues R_i and R_j $(0< i, j \leq N)$, are considered to be in contact if their spatial distance is less than a threshold *Ct*. The distance between two residues, has various definitions in the literature. It could be defined as the distance between $C_\alpha - C_\alpha$ atoms of the two residues [8], $C_\beta - C_\beta$ [10], and the minimal distance between atoms belonging to the side chain or the backbone of the residue pair [11]. In this work, the distances between $C_\beta - C_\beta$ are adopted. The distance function is [4]:

تشكر و قدرداني

- با سپاس از:
  - پروردگار يكتا
  - خانواده عزيزم
  - اساتيد گرانقدر، آقاي دكتر سرائي و آقاي دكتر كربكندي
  - اساتيد گرامي، آقاي دكتر شيخ الاسلام و آقاي دكتر موسوي
  - پروفسور Kevin Karplus
  - آقاي مهندس كاوه مهدوياني
  - تمامي معلمان و اساتيد بزرگوارم در طول دوران تحصيل
  - اعضاي آزمايشگاه داده كاوي و بيوانفورماتيک
  - دوستان خوبم
  - حضار محترم


astrojan.ini.hu