BMC Bioinformatics

BioMed Central

Protein subcellular localization prediction for Gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines

BMC Bioinformatics 2005, 6:174 doi:10.1186/1471-2105-6-174

Jiren Wang (jiren@bii.a-star.edu.sg)
Wing-Kin Sung (ksung@comp.nus.edu.sg)
Arun Krishnan (arun@bii.a-star.edu.sg)
Kuo-Bin Li (kuobin@bii.a-star.edu.sg)

ISSN 1471-2105

Article type Research article

Submission date 14 Feb 2005

Acceptance date 13 Jul 2005

Publication date 13 Jul 2005

Article URL http://www.biomedcentral.com/1471-2105/6/174

Protein subcellular localization prediction for Gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines

JirenW ang[1],W ing-KinS ung[2],A runK rishnan[1],K uo-Bin Li[1]

[1]Bioinformatics Institute,30B iopolisS treet,#07- 01M atrix,S ingapore13 8671
[2]Departmentof C omputerS cience,N ationalU niversityof S ingapore,3S cienceD rive 2,S ingapore117543

Emaila ddresses:
JW: jiren@bii.a-star.edu.sg
WS: ksung@comp.nus.edu.sg
AK: arun@bii.a-star.edu.sg
KL: kuobin@bii.a-star.edu.sg

Abstract

Background

Predictingt hes ubcellularl ocalizationof pr oteinsi si mportantf orde terminingt he functionof pr oteins.P reviousw orksf ocusedonpr edicting proteinl ocalizationi n Gram-negativeba cteriao btainedg ood results.H owever,t hesem ethodsha dr elatively lowa ccuracies fort hel ocalizationof e xtracellularpr oteins.T hispa pers tudiesw ays toi mprovet hea ccuracy forpr edictinge xtracellularl ocalizationi nG ram-negative bacteria.

Results

Weha vede velopeda s ystemf orpr edictingt hes ubcellularl ocalizationof proteinsf or Gram-negativeba cteriab asedona minoa cids ubalphabetsa nda combinationof multiples upportve ctorm achines.T her ecallof t hee xtracellulars itea nd overall recallof ou rpr edictorr each86.0% a nd89.8% , respectively,i n5- foldc ross-validation.T ot hebe stof our know ledge,t hesea ret hem osta ccurater esultsf or predictings ubcellularl ocalizationi nG ram-negativeba cteria.

Conclusions

Clustering20a minoa cidsi ntoa f ew groupsb yt hepr oposed greedy algorithm providesa ne w wayt oe xtractf eaturesf rompr oteins equencest o coverm orea djacent aminoa cidsa ndhe ncer educet hedi mensionality oft hei nputve ctorof p rotein features. Itw asobs erved thata g ood aminoa cid groupingl eadst oa ni ncreasei n predictionpe rformance. Furthermore,a p roper choiceof a s ubsetof c omplementary supportve ctorm achines constructedb ydi fferent featuresof p roteinsm aximizest he predictiona ccuracy.

Background

Subcellularl ocalizationi sa ke yf unctional attributeof a pr otein.S incec ellular functionsa reo ftenl ocalizedi ns pecificc ompartments,pr edictingt hes ubcellular localizationof unknow npr oteinsm ayb eus edt o obtainus efuli nformationa boutt heir functionsa ndt os electpr oteinsf orf urthers tudy. Moreover,s tudyingt hes ubcellular localizationof pr oteinsi sa lsohe lpfuli nunde rstandingdi seasem echanismsa ndf or developingnov eldr ugs.

Asa r esultof l arge-scale genomes equencing effortsi nr ecent years,pr oteinda taha s accumulatedi npubl icda taba nksa ta ni ncreasing rate.A nalyzingp roteind atat o extractus efulknow ledgei st huse ssentialf orpr ojectsl ikea utomatica nnotation. Iti s desirablet oha vea na utomateda ndr eliables ystemf orpr edictings ubcellular localizationof pr oteinsf roma minoa cids equences.

Anum berof e fforts[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20, 21]ha vebe enm adet op redictpr oteins ubcellularl ocalization.M ostof t hese predictionm ethodsc anb ec lassifiedi ntot woc ategories:on eba sedont he recognition ofpr oteinN -terminals ortings ignalsa ndt heot herba sedona minoa cid compositions [22].

Previousw orksha veb eenf ocusedonpr oteinl ocalizationpr edictionf orG ram-negativeba cteria.T here aref ivepr imaryl ocalizations itesi nG ram-negativeba cteria, whicha ret hec ytoplasm,t hee xtracellulars pace,t hei nnerm embrane,t heo uter membrane,a ndt hepe riplasm.P SORT I[23]i st hem ostw idelyus edt oolf or

predictingm ultiplel ocalizationsf orG ram-negativeba cteria. Itus esbi ological knowledger epresentedb y “if-then” rulesf orp redictingpr oteinl ocalizations ites. Mostof t heser ulesw ere derivedf rom experimentalobs ervations.H owever,t he PSORT Idoe snot c onsidert hee xtracellulars pace site.A dditionally,t heov erallr ecall fort heda tas et[24]onl y attains60.9% .

Gardye ta l.[24]pr esentedP SORT-Bt oi mprovet hepr edictionpe rformanceof PSORT I.P SORT-Bc ombinesi nformationof t hea minoa cidc omposition,s imilarity topr oteinsof know nl ocalization,pr esenceof a s ignalpe ptide,t ransmembranea lpha-helicesa ndm otifsc orrespondingt os pecificl ocalizationsf ora g ivenp rotein sequence,t hrougha pr obabilistica pproach. It returnsa l istof f ivepos sible localizations itesw itha ssociatedpr obabilitys cores. Ita ttainsa nov erallr ecallof 74.8%f ort hes ameda ta setm entioneda bove.

Recently, Yue ta l.[25]pr oposeda pr edictives ystemc alledC ELLOf orG ram-negativeba cteriab yus ings upportve ctorm achinesba sedonn- peptide compositions. They classified20 aminoa cidsi ntof our groups(charged,pol ar,a romatic and nonpolar)t or educet hed imensionalityo ft hei nputve ctor.F ortyS VM classifiers wereus edt opr edictt hel ocalizations ites.T heirove rallr ecallw as88.9 %. Itw asa significanti mprovementove rt hepr eviousr esultsof P SORT-B.H owever, ther ecall fore xtracelluarpr oteins wass tillr elativelyl ow at78.9% .

Thispa pers tudiesw ayst oi mprovet hea ccuracyf orpr edictinge xtracellular localizationi nt heG ram-negativeba cteria.W ee xploreda ne w wayt oe xtractf eatures frompr oteins equencesf orpr oteinl ocalizationpr edictionb y clustering20 amino

acidsi ntoa f ew groupsu singa greedy algorithm.O urm ethodf or clustering20a mino acidsc onsiderst he factorsof bot ha minoa cids'ph ysical-chemicalpr opertiesa ndt heir contextualc orrelations. Inc ontrast,t hem ethodpr esentedb yY u eta l.c lassifiest he20 aminoa cidsi nto4g roups(charged,pol ar, aromatica ndnonpol ar)b asedo nph ysical-chemicalpr opertiesof aminoa cidsa lone. Insteadof s implyc ombiningm ultiple SVMst og ivea be tterpr ediction,w epr opose as elections coref unctiona nda gr eedy algorithmt os electa s ubsetof S VMst om aximize thepr edictiona ccuracy.

Basedont hepr oposed approaches, weha vede velopeda s ystemc alledP - CLASSIFIER forpr edictingt hes ubcellularl ocalizationof G ram-negative bacteriab y usinga c ombinationof m ultiples upportve ctorm achines.T hisha sr esulted ina n improvementi nt her ecallf ore xtracellularpr oteinsf rom78.9% i nC ELLO [25] (currentlyt heb estpr edictings ystemf orG ram-negativeb acteria)t o86.0% i nP - CLASSIFIER.T heove rallr ecallof P -CLASSIFIERr eaches89.8% .T ot he bestof ourknow ledge,t hesea ret hem osta ccurater esultsf orpr edictingpr oteins ubcellular localizationi nG ram-negativeba cteria.

Results

Theda tasetus edi nt hiss tudyi sf rom[24]a ndw ase xtractedf romS WISS-PROT release40.29[26]. Itc ontains1441pr oteinsof e xperimentallyd etermined localization,w here1302 proteinsa re residenta ta singlel ocalizations itea nd139 proteinsa re residenta td uall ocalizations ites.T able11 istst henum berof pr otein sequences fromdi fferent sitesi nt heda tas et.

Thepr edictionpe rformanceof our p redictions ystemi se stimatedf roma 5- foldc ross-validationw heret he givent rainings amples are randomlypa rtitionedi nto5m utually exclusives etsof a pproximately equals izea nda pproximately equalc lassd istribution.

Iti sobs ervedt hatt herea res omepr oteins equencesi nt heda tasetc ontainingc haracter “X”.T oa voidpos siblen oisef roma mbiguousi nformation,t hepr oteine ntries containing “X”i nt hepr oteins equence aree xcludedi nt hec ross-validationt raining set,but i ncludedi nt het estings eti nt hisw ork.

Table2s howst hepr edictionr ecallf ors inglel ocalization.T her ecalli sc alculateda s TP_x/ (TP_x+ F N_x),w hereT P_xa ndF N_xr epresentt ruepos itives(numberof s amples correctly classifieda s X) andf alsen egatives(numberof s amples classified asnot X thata rea ctuallyX)ov er thepr edictives iteX .

Int hed ataset,s omep roteinsoc curi nt wodi fferents ubcellularl ocalizations.S incew e arec omparingour c ombinedc lassifierP -CLASSIFIERw itht heP -SORTBa nd CELLO classifiers,w ef ollowedt heirm ethodi ne valuatingt he classifier forpr oteins residenta tdua ll ocalizations ites,w herew ec onsidert hema spr edictedc orrectlyi f oneof t heirl ocalizations itesi spr edictedc orrectly.T able3s howst hepr ediction recallf ordua ll ocalizations.

TheM atthewsc orrelationc oefficient[27]i sus ed tom easuret hepr edictive performance forf ivep redictives ites.T heM attewsc orrelationc oefficient(*MCC*)i s definedb y:

$$MCC = \frac{(TP_x)(TN_x)-(FP_x)(FN_x)}{\sqrt{(TP_x+FN_x)(TP_x+FP_x)(TN_x+FP_x)(TN_x+FN_x)}}$$

where TP_x, TN_x, FP_x,a nd FN_xa ret ruepos itives,t ruene gatives (thenum berof samplesc orrectlypr edicteda snot X t hata rea ctuallynot X),f alsepos itives(the numberof s amplesi ncorrectlypr edicteda sX t hata rea ctuallynot X),a nd false negativesof l ocalization siteX ,r espectively. *MCC*of fersa comprehensivea ndr obust measurementf ort hepr edictivepe rformance ast hism easurementc onsidersbot h under-a ndov er-predictions.T heva lueof *MCC*e quals1f ora pe rfectpr ediction,a nd 0f ora c ompletely randoma ssignment.

Table4l istst hepe rformance comparisonsa mongP -CLASSIFIER's(ours ystem), PSORT-B's,a ndC ELLO's[25]s ystems. Ass howni nT able4,t heva luesof *MCC*of allf ives itesi nour s ystemi sg reatert hano re qualt ot heva luesi nC ELLO'ss ystem, currentlyt heb estpr edictings ystemf orG ram-negativeba cteria.M oreover,w e increaset her ecallf ort hee xtracellulars itef rom7 8.9%i nC ELLOt o86.0 %i nP -CLASSIFIER, as ignificanti mprovementf ort he extracellulars iteont hep revious results.T heove rallr ecallof P -CLASSIFIERr eaches89.8% ,w hichi sbe ttert han previousr esults.T ot heb estof our know ledge,t hesea ret hem osta ccurater esultsf or predicting Gram-negativeba cterial ocalization.

Discussion

To computationally analyse protein data, the representation of protein sequences is an important issue. A good input representation makes it easier for the SVM to identify underlying regularities andt hereforei sc rucialt ot hes uccessof S VMl earning.

Int hispa per,w ee ncode proteins equencesb yus ingt hep atternsof one aminoa cid, twoa djacent aminoa cids,t hreea djacenta minoa cids,a ndf oura djacenta minoa cids.

Ast herea re8000 and16 0000di fferentp atternsf ort het hree andf our adjacenta mino acidsc ases,c lustering20 a minoa cidsi ntos everal groupsp rovidesa w ayt or educe thenum berof uni quep atternss incei ti sdi fficultt ot raint heS VMw ithve ry large numberof f eaturess uch as160000f ora llpos siblepa tternsof foura djacenta mino acids.S incea minoa cids inpr oteinsdonot c ontributet ot hef unctionof pr oteins independentlya ndf unctionalpa tternsi npr oteins aree mbeddeda ss equence correlations,a minoa cidsm aynot be groupedba sedont heirph ysical-chemical propertiesa lone[28].F ort hepr edictiont ask,a gooda minoa cid groupingl eadst oa n improvementi npr edictionpe rformance.

Iti sobs ervedt hatt hepr edictionr esultsf romS VMsc onstructedb ydi fferentl engths ofa djacenta minoa cidp atterns,e .g.t hep atterns ofa s inglea minoa cid anda mino acidpa irs,a rec omplementary. Thati s,t here ares omec asesw heret hepr ediction madeb yt heS VM constructedb ypa tternsof s omepa rticularl engthi sc orrectw hile thepr edictionm adeb yt heS VMc onstructedb yp atternsof a notherl engthi s incorrect,a ndvi cev ersa.T herefore, combiningc omplementaryr esultspr ovidesa wayt oi mprovet hepr edictiona ccuracy.H owever,c ombininga llc omplementary resultst ogetherm aynot bea goodc hoice.T herefore,w ep roposet oc hoosea s ubset ofc omplementarys upportve ctorm achinespr operlyt hatw illm aximizet hepr ediction accuracy.

Aftera nalyzingt hepr edictiver esults,i ti sobs ervedt hatt herea res omepr otein sequencest hatc annotbe predictedc orrectlyb y any SVMi nt hec ombined classifier. Itm eanst hatt hesepr oteins equences cannotbe correctlyc lassifiedb yt heir

composition.T hisi st her easonw hyt her ecallof s omepr edictives itesi nG ram-negativeba cteria cannot bef urtheri mproved.

Sincew ea rec omparing ourc ombinedc lassifierP -CLASSIFIERw itht he P-SORTB andC ELLOc lassifiers,w eus et hes ameda tas et ast heirs.W edi dnot c heckt he sequence redundancyi nt heda taset.A st hel evelo fs equencer edundancyn ormally stronglya ffectspr edictiona ccuracy, removingt hosepr oteins equencesw hichha ve highs equencei dentity(e.g.m oret han40%) withe achot heri nt heda tasetc ana void redundancy andbi as.

Insteadof givingf ullc reditf ordua l-localizedpr oteinsi fe itherof t hes ites ispr edicted correctly, wea lsoe valuatet hepr edictionpe rformanceb yc ounting“ half”c orrect whenonl yon eof t hes itesof dua l-localizedpr oteinsi spr edictedc orrectly. Table5 showst heirpr edictionr ecalls.T hef ullc reditf ord ual-localizedpr oteinsi sonl y given whent wopos siblel ocalizations itesw itht het opt woa ssociatedp robabilitys cores matcht oa ctualdua ll ocalizationsof t hepr otein.T hec orrespondingove rallr ecallf or predictingdu all ocalizationsonl yr eaches67.3% . Topr operlyde alw iths ubcellular localizationsf orpr oteinsr esidenti ns everaldi fferents itesi sa c hallengingp roblem. Thepa per[5]a ddressedt hepr oblemof s ubcellularl ocalizationsf orpr oteinsr esident ins everaldi fferents ites.

There aret hreem ethods usedf orc ross-validation test:t hei ndependentda tasett est,n-foldc ross-validationt est,a ndt hel eaveone out c ross-validationt est.A mongt hese methods,t hel eaveone o utc ross-validationt esti st hem ostr igorousa ndobj ective[29, 42].H owever,t hel eave oneout c rossva lidationt esti sve rye xpensive

computationallya ndi sof teni mpracticalf orl arge datasets.T hen- foldc ross-validationt estpr ovidesa bi as-freee stimationof t hea ccuracy[30]a tm uchr educed computationalc osta ndi sc onsidered asa na cceptablet estf ore valuatingp redictive performanceo fa na lgorithm[31]f orl argeda tasets.

Conclusions

Thispa peri ntroducesa p roteins ubcellularl ocalizationpr edictionm ethodus ing aminoa cids ubalphabets anda c ombinationof m ultiples upportve ctorm achines. Them ainc ontributionsof our w orki nclude:(1)A ne ww ayt oe xtractf eaturesf rom proteins equencesb yc lustering20 aminoa cidsi ntoa f ew groupsus ingt hepr oposed greedya lgorithmt or educet hei nputdi mensionalityof s upportv ectorm achines.O ur methodf orc lustering20 aminoa cidsc onsidersno tonl yt hef actorof t he aminoa cids' physical-chemicalpr opertiesbut a lsot hef actorof theirc ontextualc orrelations.(2)A selections core functiona nda gr eedy algorithma repr oposedt os electa s ubsetof candidates upportv ector machinest om aximizet hec ross-validationa ccuracyi nstead ofs implyc ombiningm ultiples upportve ctorm achinest og ivebe tterpr ediction.(3)A web-baseds ystemha sbe ende veloped forpr edictingpr oteins ubcellularl ocalization ofG ram-negativeba cteria. It allowspe oplet os ubmitm ultipleG ram-negative bacteriap roteins equencest ope rformpr oteins ubcellularl ocalizationpr ediction. Iti s availablea t[43].

Clustering20a minoa cidsi ntoa f ew groupsb you rpr oposed greedya lgorithm providesa ne w wayt oe xtractf eaturest oc overm orea djacenta minoa cidsf rom proteins equencesa ndr educet hedi mensionality oft hesef eatures.S ince aminoa cids

inpr oteinsdonot c ontributet ot hef unctionof pr oteinsi ndependently,i tm aynot b ea goodi deat ogr oupa minoa cidsba sedont heirph ysical-chemicalpr opertiesa lone.F or thepr edictiont ask,a gooda minoa cid groupingl eadst oa ni ncreasei npr ediction performance. Furthermore,pr operly choosing as ubsetof c omplementarys upport vectorm achinesc onstructedb ydi fferentf eatures ofpr oteinsm aximizest he predictiona ccuracy.

Methods

Support vector machines

SupportV ectorM achines(SVMs)ha vebe enw idelyus edi nt he analysiso fbi ological data[32,33,34] .S VMi sa r elativelyne wf amily ofl earningm ethodsa nd hass ome theoreticals upportf roms tatisticall earningt heory[35,36] . SVMnon- linearlym aps thei nputs pacei ntoa hi ghdi mensionalf eatures pace, ands eeks ah yperplanei nt his spacet hats eparatest hep ositives amplesf romt hene gativeone sw itht hel argest possiblem argina ndopt imizest het rade-offb etweeng ood classificationa ndl arge margin. Insteadof explicitlym appingt heobj ectst ot hehi ghdi mensionalf eature space,S VMus uallyw orksi mplicitlyi nt hef eatures paceb yonl y computingt he correspondingk ernelbe tweena nyt woobj ects.

Severalpa rametersne ed tobe s etdur ingt heS VMt rainingpha se.T hesepa rameters includet her egularizationpa rameter,w hich controlst het rade-offbe tween good classificationa ndl arge margin,t hek ernelt ype,a ndt heke rnelpa rameters.T hese parametersa ret unedba sedont hec riteriaof c ross-validationa ccuracy. Ther adial basisf unction(RBF)ke rneli sus edf ora llour e xperimentsa ndt hes oftwareB SVM [44],a m ulti-classS VM[37], isus edi nt hisw ork.

Protein features

Thea minoa cidc ompositionsi nt hef ullor pa rtials equences are considereda sg lobal features,w hichr epresentt heove ralls imilarity amongm ultiplepr oteins equences. In thispa per,t he globalf eaturesa reus eda st hei nputof t heS VMst opr edict protein subcellularl ocalization.

a. W-gram protein encoding

Twot ypeso ff eatures arec onsideredi nour work: W-gram and gapped2 -gram.A W -grami sde fineda sp atternsof W (W ≥1) c onsecutivea minoa cidr esidues withouta ny gaps anda gapped2- grami sde fineda st woa minoa cidr esidues withs omes pecified numberof gapsi na p roteins equence.H ere,a gapped2 -grami sa lsor eferredt oa sa 2-gram.T hem ainpu rposeof i ntroducingt he gappede ncodingf eaturesf or 2-grami s toi ncreaset henum berof 2- gramf eature candidates.

Fore achpr oteins equenceP a nde achW -gram(orf eature)F ,l etN (P,F)be t he numberof oc currenceso fF i nt hepr oteins equenceP .F urther,l etT (P,W)be t het otal numberof pos sibleW -gramsi nP , *length*(P)be t hel engthof P ,a ndG (F)b et he specifiednum bero f gaps.W eha veT (P,W)= *length*(P) –W + 1 –G (F), whereG (F) =0i fW ≠2a ndG (F) ≥ 0i fW = 2.T hef eaturev alueU (P,F)w ithr espectt ot he featureF andt hes equenceP i sde fineda sN (P, F) /T (P,W).F ore xample,s upposeP =“ LAEVLAAA”a nd F =“ LA” (withouta ny gaps),t hent hef eaturev alue U(P,F)i s 2/ (8–2+ 1–0) = 0.28 57,w here F= “LA”, N(P,F)= 2, *length*(P)= 8,W =2,G (F)= 0,a ndT (P,W)= 7. Intuitively,U (P, F)m easurest hepr oportionaloc currencesof F among allpos sibleW -gramsi nP .T hism easurementi sl engthi ndependent.

Int heW -grampr oteine ncodingm ethod,t het otal numberof di fferentpos sible featuresi s20 ^{w}.

b. Amino acid subalphabets

Iti sdi fficultt ot raint he SVMw ithve ryl argenu mberof f eaturess ucha s 8000f or3 -gram.T or educedi mensionality,one w ayi st oc lassifyt he20 aminoa cidsi ntos mall numberof groupsba sed ont heirph ysical-chemicalpr operties.A llm embersi nt he samegr oupc anbe r epresentedb yone s ymbol.T hem erged aminoa cida lphabetha s fewert han20s ymbolsa ndi sc alledt hea minoa cids ubalphabet,w hichc anbe us edt o re-encodet heo riginalpr oteins equences.T her e-encodedpr oteins equencesha ve fewerf eatures. Fore xample,i ft henum berof s ymbolsi na na lphabeti sr educedf rom 20t o6,t henum berof 3- gramf eaturesi sr educed from4000(20× 20 ×20)t o216(6 ×6× 6) .R educingt hen umberof f eaturest oa m anageables izef orS VMsc anhe lpt o improvet hepr edictivep erformance.

Thispa pers uggestsopt imizingt heg roupingb yu singt hepr oposed greedya lgorithm, whichc onsiderst he factorsof bot ht hea minoa cids'ph ysical-chemicalpr opertiesa nd theirc ontextualc orrelations,i nsteadof us ingt he groupingb asedont heirp hysical-chemicalpr opertiesa lone.N otet hatt herea rea n exponentialnum berof w ayst o group the20a minoa cids.F ore xample,t herea re580606 446a nd45232115901w ayst o divide20a minoa cidsi nto3a nd4gr oups,r espectively.T henum berof s ubalphabets withm g roups($1 \leq m \leq 20$)f ort hepr oteina lphabets izeof 20, $N(m)$c anb ec alculated byt hef ormula[28]be low.

$$N(m) = \begin{cases} 1 & i \quad fm = 1 \\ \frac{m^{20}}{m!} - \sum_{k=1}^{m-1} \frac{N(m-k)}{k!} & if 1 < m \leq 20 \end{cases}$$

We learn the local optimal grouping based on a greedy algorithm using the SVM classification algorithm to evaluate the fitness of each candidate subalphabet, where the criteria for evaluation is the 5-fold cross-validation accuracy.

c. Search for amino acid subalphabets

This section presents our greedy algorithm for finding a good grouping for the amino acids. Given a particular subalphabet encoding schema S, supposing N_g and T_c are the predefined number of groups and threshold of cross-validation accuracy, respectively. Further, we assume the parameters of a SVM to evaluate the fitness of a candidate subalphabet are given. These SVM parameters can be set either by the values suggested by the SVM software or by the tuning result of the SVM, which is constructed from features re-encoded by grouping 20 amino acids based on their physical-chemical properties, according to the criteria of cross-validation accuracy. For a particular subalphabet encoding schema S, let the grouping score h(S) be the cross-validation accuracy when prediction is done by a SVM using W-gram and the subalphabet scheme S. h(S) can be used to measure the goodness of the grouping S.

Table 6 shows an example of clustering 20 amino acids into 4 groups for the 4-gram protein encoding method using the proposed greedy algorithm. The initial node with 4-group assignment is set to {(A, G, I, L, M, P, V), (C, N, Q, S, T), (D, E, K, H, R), (F, W, Y)}, which is based on the physical-chemical property of amino acids. The

process for searching for an amino acid subalphabet is iterated until it reaches a local maximal grouping score at 79.0285%, where the final four groups are {(I, L, M, V), (N, S, T), (C, D, E, H, K, Q, R, Y), (A, F, G, P, W)}. Note that some group members in the classified result have the same physical-chemical property of amino acids. For example, the amino acids A, F, G and W in the fourth group (A, F, G, P, W) are all hydrophobic. In particular, the amino acids F and W are aromatic while amino acids A and G are tiny. Further, the hydrophilicity scale indices of A, G, P, and W have approximately the same values in the amino acid index database [38], which suggests that the hydrophilicity of amino acids may be an important property in classifying the 20a minoa cids.

The proposed greedy algorithm to search for amino acid subalphabets is described in Table 7. The greedy local search [39] has been used for learning the subalphabets. In the search tree [39], every node represents an amino acid subalphabet encoding schema. The child nodes of a node are subalphabets encoding schemata, which are generated by moving every group member to each other group if the number of membersi nt hisg roupi s greatert hanone .

Thisa lgorithmi sc omposedof t hef ollowingf ours teps.F irst,t he20a mino acidsa re initiallydi videdi ntoN $_g$ groupse itherr andomlyw itha pproximatelyt hes ames izeor basedons omeph ysical-chemicalpr opertiesof t he20a minoa cids.A minoa cidsi nt he samegr oupa rede noted byone s ymboli na s ubalphabet.S upposet hec urrent subalphabete ncodings chemai sr epresentedb y currentnode ,i ts grouping scorei s calculated wheret he groupings corei st hec ross-validationa ccuracyw hen prediction isdone b y aS VMus ing W-gram andt hiss ubalphabets cheme.

Second,a llc hildnode so ft hec urrentnode a re generated. Ift herei sonl yo nem ember ins omeg roup,t hism emberc annotm ovet o anyo thergr oup.O therwise,t het otal numberof groupsw illbe l esst hanN $_g$.T herea rea tm ost20× (N_g–1) pos siblec hild nodesi nt hes earchings paces incet herea re20a minoa cidsa nde acha minoa cidc an onlym ovet oa tm ost(N_g –1) ot her groups. Ift he highest groupings corea mongt he childnode si sg reatert hant heg roupings coreo ft hec urrentnode ,t hisc hildnode w ill becomet he currentnode .

Third,t hea bovepr ocess fors earchingt hec hildno dew itht hehi ghestgr oupings core among allc hildnode sw illbe r epeatedunt ilt heg roupings coresof a llc hildnode sa re lesst hant hegr oupings coreof t he currentnode .

Fourth,i ft he groupings corei nt hef inalc urrentno dei sg reatert hanT $_c$,t he N_g groups int hec urrentnod ew illbe comet he acceptedm ergeds ubalphabets.O therwise,w e randomly re-generatet hec urrentnod ea ndr epeatt heS teps2t o4a bove.

Thet rainings equencesa redi videdi ntot wopa rts:O nepa rti sus edf orc hoosingt he subalphabetw hilet heot heri sus edf or evaluatingt hepe rformanceo fa s ubalphabet.

Thegr eedya lgorithmi sa ppliedt or educet henu mberof W -gramf eatures. In particular,f or3- gram,w ec lassifyt he20a minoa cidsi nto6,7,a nd8g roups.F or4-gram,w ec lassifyt he20 aminoa cidsi nto4g roups.T henum berof featuresi s m^W, where mi st henum berof g roups andW i st henum berof pr oteinpe ptides inW -gram

encodingm ethods.F or example,t henum berof f eaturesi s6× 6× 6 =216 for6 groupsi n3- grame ncodingm ethod.

Multiple SVMs

Duet ot hena tureo ft he multi-classc lassification,i tm aynot be e asyt oobt aina s ingle SVMt hatc anr eturnhi gha ccuracies fort hes ubcellularl ocalizationpr ediction. Therefore,m ultipleS VMsa ret rainedf romdi fferentf eatures andt heirr esultsa re combinedus ingvot ing.

Currentlym ostof t he existingpr oteins ubcellularl ocalizationpr edictions ystems usingS VMsonl yus et hef eaturesge neratedf rom 1-gramor 2- grampr oteine ncoding methods.F ore xample,t hee xtractedf eatureso fa minoa cidc ompositions[2]a nd featuresof a minoa cidpa ira ndga pped aminoa cidpa irc ompositions[40]c anbe considered ast hef eaturesg eneratedf romt he1- grama nd2 -grame ncoding methods, respectively.

Asm any functionalpa tternsi npr oteinsa re embeddeda ss equence correlations,i ti s expectedt hatm orei nformationw illbe i ncludedby combiningc lassifiersc onstructed fromf eatures generated by1- gram,2- gram,3- gram,a nd4 -grampr otein encoding methods,i nsteadof j ustus ingt hec lassifiersc onstructedf rom1- grama nd 2-gram encodingm ethodss ince morea djacenta minoa cidr esiduesw illbe c onsidered.

Int hispa per,t hef ollowingf ourt ypesof f eatures aree xtractedf rompr otein sequences.T hef irstt ype ist he1- grame ncoding feature,w hichi ncludesa minoa cid compositionsa ndt hepa rtitioneda minoa cidc ompositions,w heret hepr otein

sequencei sp artitionedi ntoP pa rtsw itha pproximatelyt hes amel ength.T het otal numberof t hese features is20× P . Int hisw ork,P iss etf rom2t o6.T hes econdone is2- grame ncodingf eature,w hichi ncludes aminoa cidpa ira nd gapped aminoa cid pairc ompositions,w heret henum berof f eaturesi s400(20× 20) a ndt henu mberof gapsi ss etf rom1t o2.T hepur poseof i ntroducingt he gappede ncodingf eaturesonl y for2- grami st oi ncrease thenum berof 2- gramf eaturec andidates.T het hirdone i st he 3-grame ncoding feature,w heret he20 aminoa cidsa redi videdi nto6,7,a nd8g roups whosenum bersof featuresa re216 (6× 6 ×6) ,34 3(7× 7× 7) , and512(8 ×8× 8) , respectively.T hel astone i st he4- grame ncoding method,w heret he20 aminoa cids aredi videdi nto4 groups,w hosenum berof featuresi s256(4× 4× 4 ×4) .

Feature selection

Wea pplyt hew rappera pproach[41]i nt heba ckwarde liminationve rsiont os electt he features ubsetf orou rS VMc lassifiersa ndus e5- foldc ross-validationa ccuracya st he criteriaf ore valuation.

LetS VM_a andS VM_b bet heS VMc lassifiersus inga llf eaturesa ndf eatures selectedb y thew rappera pproach, respectively. Althought he predictiona ccuracyof S VM_b is improved,t hepr ediction resultsf romS VM_a and SVM_ba redi fferent.T here ares ome casesw heret hep redictionm adeb yS VM_a isc orrectw hilet hepr edictionm adeb y SVM_bi snot c orrect,a nd viceve rsa.T herefore,bo thS VM_aa ndS VM_bc an be considered asc andidates tobui ldt hef inalc ombinedc lassifier.

SVM subset selection

DifferentS VMsg ivedi fferentpr edictions.O new ayt o combinet heirpr edictionsi sb y voting.T hati s,e achpr oteins equencei sa ssigned toa c lassw itht hem ostvot es.F or casesw heret woo rm ore classes gett hem ostvot es,w ea ssignt hesec asest ot he predictiver esultsb yone oft heS VMs,w hich gets them ostnum berof c orrect predictionsf ora llt hese cases.

SupposeS i sa s etof pr oteins equences,N i st hen umberof c andidateS VMs,M = {SVM_1,S VM_2,…,S VM_N}i st hes etof c andidate SVMsde finedpr eviously,V_1(S, M)i st henum berof c orrectpr edictionsc lassified byM w ithonl yone c lass correspondingt ot hem ostvot e,a ndV $_2$(S,M)i st henum berof t he correctp redictions byt he assignedS VMw hent woor m ore classesc orrespondt ot hem ostvot e.T he selections core function V(S,M)i sde fineda sV_1(S,M)+ V $_2$(S,M)a ndi sus edt o selecta s ubsetof allc andidateS VMst of orma c ombinedc lassifier,w hich maximizes thec ross-validationa ccuracy.T hepr oposed greedy algorithmt os electa s ubsetof M isde scribedi nT able8.

Thisg reedya lgorithmc onsistsof t hef ollowingt wos teps.F irst,s etM = {S VM_1, SVM_2,…,S VM_N},S $core_{max}$= V (S,M),S et_{max}= M,a ndi = N - 1.S econd,f ore very memberS VM_r ∈M (1 ≤ r ≤N),r emoveS VM_rf romM a ndc alculatet heva lueof i ts correspondings elections coref unctionV (S,M –{S VM_r})(1 ≤r ≤N).S upposef or someS VM_j(1 ≤j ≤N),V (S,M –{S VM_j})i se qualt oV_{max},t hem aximal valueof a ll V(S,M –{S VM_r})(1 ≤r ≤N),t henupda tet he following:M = M –{S VM_j},S $core_{max}$ =V $_{max}$,S et_{max}= M ,a ndi = i –1.T hepr ocessf or removings omeS VM_p (1 ≤p ≤N) willc ontinueunt ili = 1,t hati s,onl yone S VMi sl eft.T henS et_{max}i ss electedt obe t he combinedc lassifier.

Wec anus et hepr edictionr esultsof f our-fiftht rainingpr oteins equencest os electa subsetof S VMsa ndus et hepr edictionr esultsof t her estof one -fiftht rainingpr otein sequencest o evaluatet hepe rformanceo ft her esultof t heS VMs ubsets election.

Int hisw ork,15S VMsa res elected andc ombined tof ormt hef inalc lassifier.T able9 showst hee ncodingm ethodsof i nputve ctorsi nt hef ifteens electedS VMs. Rows12, 13,a nd14r epresent3di fferentm ergeds ubalphabets,w hicha re{(A, F,G , P,W),(C, D,E ,H ,K ,Q ,R ,Y), (N,S ,T),(I, L,M ,V)},{(A,C ,M ,P ,V),(F, I, L,W),(D,E ,H , Q,R),(G,K ,N ,S ,T ,Y)},a nd{(A,G ,P ,Q ,Y),(C,D ,E ,H ,K ,M ,R),(N,S ,T),(F, I, L,M , V)},r espectively.R ows4,7a nd15r epresentt hes amee ncodingm ethoda s ther ows3,6a nd14but withf eatures election.

Weha vec onducteds omee xperimentsonc onstructingS VMsb yus ing5 -gram encodingm ethod.P reliminary experimentalr esultss howt hatt hec ross-validation accuraciesp redictedb yS VMc onstructedb y3- gram,4- gram,a nd5 -grame ncoding methodsa renot s atisfactory whent henum berof groupsi sl esst han6,4, and4, respectively.W henw ei ncreaset henum bero f groupst o4f or5- gram,t het ime requiredt ot raint he correspondingS VMa ndc alculatet he5- foldc rossva lidation accuracyi sr elativelys lowa st henum bero ff eaturesr eaches1024 (4× 4 ×4× 4× 4) . Therefore,onl y1- gram, 2-gram,3- gram,a nd4- grame ncodingm ethodsa re consideredi nt hispa per. Furthermore,t he20a minoa cidsa rec lassifiedi nto6,7,a nd 8g roups for3- grama nd 4g roups for4 -grame ncodingm ethods,r espectively.

Sincet herea ret oom any zeroe lementsi nt hee ncodingr esults,2- gram,3- gram,a nd 4-grampr otein's encodingm ethodsa renot a ppliedt ot hosec asesw heret hepr otein sequences arep artitionedi ntoP (P> 1) pa rtsw itha pproximatelys amel ength.

Authors' contributions

JWde velopedt hem ethods,bui ltt hes ystem andd raftedt hem anuscript.W S,A Ka nd KLp articipatedi ns ystemde sign,pr ovidedv aluablec omments,a ndhe lpedt odr aft them anuscript.

Acknowledgements

Thea uthorsw ouldl iket ot hankt hea nonymousr eviewers whosec ommentsha ve helpedus i mprovet hem anuscript.

References

1. Emanuelsson,O .,N ielsen,H .,B runak,S .,vonH eijne,G .: **Predicting subcellularl ocalization ofp roteinsb asedon t heirN -terminala minoac id sequence.** *JM olB iol.*,2000,300(4):1005 -1016
2. Hua,S .a ndS un,Z .: **Supportve ctor machineap proachf orp rotein subcellularl ocalization prediction.** *Bioinformatics*,2001,17(8):721 -728
3. Horton,P .a ndN akai,K . **Better Predictionof P roteinC ellularL ocalization Sitesw itht he *k*N earestN eighborsC lassifier.** *Intellig.Sy st.M ol.B iol.* 1997,5: 147- 152
4. Nakashima,H . andN ishikawa,K .: **Discriminationof I ntracellularan d Extracellular ProteinsU singA minoA cidC ompositionan dR esidue-pair Frequencies**. *JM olB iol.,*1994,238(1):54 -61
5. Cai,Y .D.a ndC hou,K .C.: **Predicting22p roteinl ocalizationsi nb udding yeast**. *BiochemicalandB iophysicalR esearchC ommunications*,2004,323: 425-428
6. Chou,K .C.: **Prediction ofp roteinc ellularat tributesu sing pseudo-amino acidc omposition.** *PROTEINS:St ructure,F unction,andG enetics*,2001, 43: 246-255(Erratum:i bid.,2001,44: 60)
7. Chou,K .C.a ndC ai,Y .D.: **An ewh ybridap proacht op redicts ubcellular localizationof p roteins byi ncorporatingG ene ontology**. *Biochemicala nd BiophysicalR esearchC ommunications*,2003,311: 743 -747
8. Chou,K .C.a ndC ai,Y .D.: **Predictionan dc lassificationof p rotein subcellularl ocalization:s equence-order effecta ndp seudoam inoac id**

composition. *Journalof C ellularB iochemistry*,20 03,90: 1250- 1260 (Addendum,i bid.2004,91(5):1085)

9. Chou,K .C.a ndC ai,Y .D.: **Predictings ubcellularl ocalizationof p roteins by hybridizing functional domainc ompositionan dp seudo-aminoac id composition**. *Journalof C ellularB iochemistry*,2004,91: 1197-1203
10. C hou,K .C.a ndC ai,Y .D.: **Predictionof p roteins ubcellularl ocationsb y GO-FunD-PseAAp redictor.** *BiochemicalandB iophysicalR esearch Communications,*2004,1236- 1239.
11. F eng, Z.P.: **Predictionof t hes ubcellularl ocationof p rokaryoticp roteins basedon an ewr epresentationof t heam inoac idc omposition**. *Biopolymers*,2001,58: 491 -499
12. F eng, Z.P.a nd Zhang,C .T.: **Predictionof m embranep roteint ypesb ased ont heh ydrophobici ndexof am inoac ids**. *Journalof P roteinC hemistry*, 2000,19: 269 -275
13. F eng, Z.P.a nd Zhang,C .T.: **Predictionof t hes ubcellularl ocationof prokaryoticp roteinsb asedon t heh ydrophobicityi ndexof am inoac ids**. *IntJ B iolM acromol*,200 1,28: 255 -261
14. P an,Y .X ., Zhang, Z. Z., Guo,Z .M ., Feng,G .Y ., Huang, Z.D .a nd He, L.: **Applicationo fp seudoa minoac idc ompositionf orp redictingp rotein subcellularl ocation:s tochastics ignalp rocessingap proach**, *Journalof ProteinC hemistry*,2003,22: 395- 402
15. W ang,M .,Y ang,J ., Liu, G.P.,X u,Z .J .a ndC hou,K .C .: **Weighted-support vector machinesf orp redictingm embranep roteint ypesb asedon p seudo aminoac idc omposition**. *ProteinE ngineering,D esign,andSe lection*,200 4, 17:509- 516
16. W ang,M .,Y ang,J .,X u,Z .J .a ndC hou,K .C .: **SLLE forp redicting membranep roteint ypes**, *Journalof T heoreticalB iology*,2004,232: 7 -15
17. X iao,X .,S hao,S .,D ing, Y.,H uang, Z.,H uang,Y .a ndC hou,K .C .: **Using complexity measuref actort op redictp roteins ubcellularl ocation**. *Amino Acids*,2005,28(1):57- 61
18. Yuan, Z.: **Predictionof p roteins ubcellularl ocationsu singM arkovc hain models**. *FEBSL etters*,1 999,451,23 -26
19. Zhou,G .P .: **Ani ntriguingc ontroversyove rp roteins tructuralc lass prediction.** *Journalof P roteinC hemistry*, 1998,1 7:729 -738
20. Z hou,G .P . &A ssa-Munt,N .: **Somei nsightsi ntop roteins tructuralc lass prediction.** *PROTEINS: Structure,F unction,andG enetics*,2001,44: 57 -59
21. Z hou,G .P.a nd Doctor, K.: **Subcellularl ocationp redictionof ap optosis proteins**. *PROTEINS:St ructure,F unction,andG enetics*,2003,50: 44 -48
22. N akai,K .: **Proteins ortings ignalsa ndp redictionof s ubcellular localization**. *Adv.P roteinC hem.*,2000,54: 277- 344
23. N akai,K .a nd Kanehisa, M.: **Experts ystemf or predictingp rotein localizations itesi nG ram-negativeb acteria.** *Proteins.*1991,11(2):95 -110
24. J ennifer L.G ardy,C ory Spencer,K eW ang,M artinE ster,G aborE .T usnady, IstvanS imon,S ujunH ua,K atalinde Fays,C hristophe Lambert,K entaN akai andF ionaS .L. Brinkman. **PSORT-B:i mprovingp roteins ubcellular localizationp redictionf orG ram-negativeb acteria**. *NucleicA cids Research.*2003,31: 361 3-17.
25. Y u,C .-S., Lin,C .-J.,a ndH wang,J .-K.: **Predictings ubcellularl ocalization ofp roteinsf orG ram-negativeb acteriab ys upportve ctor machinesb ased onn -peptidec ompositions.** *ProteinSc ience.*200 4,13(5):1402 -1406

26. Bairoch,A .a ndA pweiler,R .: **TheS WISS-PROTp roteins equence databasean di tss upplementT rEMBLi n2000.** *NucleicA cidsR esearch.* 2000,28: 45- 48

27. Matthews,B .W., **Comparisonof t hep redictedan dob serveds econdary structureof T 4p hagel ysozyme**. *Biochim.B iophys.A cta*.,1975,405(2): 442-451

28. Andersen,C .A.F.a nd Brunak,S . **Representationof P rotein-Sequence Informationb yA mino AcidS ubalphabets**. *AIM agazine.*2004,25(1):9 7-104

29.M ardiaK .V.,K entJ .T.,B ibbyJ .M.: **MultivariateA nalysis**. London: AcademicP ress,1979,3 22-381

30.S toneM : **Cross-validatoryc hoicean das sessmentof s tatisticalp redictions**. Journalof t heR oyalS tatisticalS ociety1974,36: 1 11–147.

31. KohaviR : **Wrappersf orp erformance enhancementan dob liviousd ecision graphs**.P hDt hesis,S tanfordU niversity1995.

32. Brown,M .P .S .,G rundyW .N ., Lin,D .,C ristianini,N .,S ugnet,C .W .,F urey, T.S .,A res,M .,a ndH ausslerD .: **Knowledge-basedan alysiso fm icroarray genee xpressiond atab yu sings upportve ctor machines.** *PNAS*.2000,97: 262-267

33. Lee,Y .a nd Lee,C .-K.: **Classificationo fm ultiplec ancert ypesb y multicategorys upport vector machinesu singg enee xpressiond ata**. *Bioinformatics*,2003,19 **:** 1132-1139

34.W ard,J . J.,M cGuffin, L.J .,B uxton,B .F .,a ndJ ones,D .T .: **Secondary structurep redictionw ith supportve ctor machines.** *Bioinformatics*,200 3, 19:1650 -1655

35.V apnik,V .: **Then atureof s tatisticall earningt heory**.S pringer-Verlag,N ew York,1995

36.V apnik,V ., **Statisticall earningt heory**,J ohn-Wiley, NewY ork,1998

37.H su,C .- W.a nd Lin,C . –J., **As impled ecompositionm ethodf ors upport vector machines**. *MachineL earning*,2002,46: 291- 314.

38.K awashima,S .a ndK anehisa,M .: **AAindex:am inoac idi ndexd atabase.** *NucleicA cidsR es.*2000, 28:374

39.R ussel,S .J.a ndN orvig, P.: **Artificiali ntelligence:a modernap proach**, PrenticeH all.2003

40.P ark,K .-J.a ndK anehisa,M .: **Predictionof p roteins ubcellularl ocationsb y supportve ctor machinesu singc ompositionsof aminoac idsan dam ino acidp airs**. *Bioinformatics*,2003,19(13):1656 -1663

41.K ohavi,R .a ndJ ohn,G .H.: **Wrappersf orf eatures ubsets election**. *Artificial intelligence*.1997,97(1-2):273- 324

42.C hou,K .C .& Zhang,C .T . **Review: Prediction ofp roteins tructural classes**.*C riticalR eviewsi nB iochemistryandM olecularB iology,1995,* 30, 275-349

43. **Proteins ubcellularl ocalizationp redictionf or Gram-negativeb acteria** [http://protein.bii.a-star.edu.sg/localization/gram-negative/].

44. **BSVM**[http://www.csie.ntu.edu.tw/~cjlin/bsvm/index.html].

Tables

Table 1 - Number of protein sequences in different sites

Localizations ites	No.
cytoplasmic	248
inner membrane	268
periplasmic	244
outmembrane	352
extracellular	190
cytoplasmic/ i nnerm embrane	14
membrane/ pe riplasmic	49
outerm embrane/ e xtracellular	76
Alls ites	1441

Table 2 - Prediction recall for a single localization.

Localization	Recall($TP_x/(TP_x+FN_x)$)
Cytoplasmic	94.8%(235/ 248)
Extracellular	83.2%(158/ 190)
Innermembrane	88.1%(236/ 268)
Outermembrane	93.2%(328/ 352)
Periplasmic	86.9%(212/ 244)
Overallr ecall	89.8%(1169/1302)

Table 3 - Prediction recall for dual localizations.

Localization	Recall ($TP_x/(TP_x+FN_x)$)
Cytoplasmic/ i nnermembrane	92.9%(13/14)
Outermembrane/ e xtracellular	98.9%(75/76)
Periplasmic/ i nnermembrane	75.5%(37/49)
Overallr ecall	89.9%(125/139)

Table 4 - Performance comparisons among P-CLASSIFIER's, PSORT-B's, and CELLO's methods.

	P-CLASSIFIER		CELLO		PSORT-B	
Localization	**Recall**	***MCC***	**Recall**	***MCC***	**Recall**	***MCC***
Cytoplasmic	94.6%	0.85	90.7%	0.85	69.4%	0.79
Extracellular	86.0%	0.89	78.9%	0.82	70.0%	0.79
Innermembrane	87.1%	0.92	88.4%	0.92	78.7%	0.85
Outermembrane	93.6%	0.90	94.6%	0.90	90.3%	0.93
Periplasmic	85.9%	0.81	86.9%	0.80	57.6%	0.69
Overallr ecall	89.8%	-	88.9%	-	74.8%	-

Table 5 - Prediction recall for dual localizations when "half" predictions are only counted as half correct.

Localization	Recall ($TP_x/(TP_x+FN_x)$)
Cytoplasmic/ i nnermembrane	75.0%(10.5/14)
Outermembrane/ e xtracellular	84.2%(64/76)
Periplasmic/ i nnermembrane	38.8%(19/49)
Overallr ecall	67.3%(93.5/139)

Table 6 - An example of clustering 20 amino acids into 4 groups.

Searchings tates	Cross-validation accuracy	Actions
(A,G , I, L,M ,P ,V) (C,N ,Q ,S ,T) (D,E ,H ,K ,R) (F,W ,Y)	71.2413%	Move' G'f rom group 1t og roup4
(A, I, L,M ,P ,V) (C,N ,Q ,S ,T) (D,E ,H ,K ,R) (F,G ,W ,Y)	74.0941%	Move' A'f rom group 1t og roup4
(I, L,M ,P ,V) (C,N ,Q ,S ,T) (D,E ,H ,K ,R) (A,F ,G ,W ,Y)	75.9445%	Move' P'f rom group 1t og roup4
(I, L,M ,V) (C,N ,Q ,S ,T) (D,E ,H ,K ,R) (A,F ,G ,P ,W ,Y)	77.5636%	Move' C'f rom group 2t og roup3
(I, L,M ,V) (N,Q ,S ,T) (C,D ,E ,H ,K ,R) (A,F ,G ,P ,W ,Y)	78.4888%	Move' Q'f rom group 2t og roup3
(I, L,M ,V) (N,S ,T) (C,D ,E ,H ,K ,Q ,R) (A,F ,G ,P ,W ,Y)	78.9514%	Move' Y'f rom group 4t og roup3
(I, L,M ,V) (N,S ,T) (C,D ,E ,H ,K ,Q ,R ,Y) (A,F ,G ,P ,W)	79.0285%	Reachl ocalm aximal groupings core and stop.

Table 7 - Algorithm for amino acid subalphabets searching

1	current_node ← the initial group assignment by dividing the 20 amino acids into N_g groups.
2	REPEAT
3	be st_node ←c urrent_node
4	R EPEAT
5	c urrent_node ←be st_node
6	g eneratea llc hild nodesof t hec urrentnode i nt hes earcht ree.
7	be st_node ←t hec hildnode w itht hehi ghest *h*-valuea monga llc hild nodesof t hec urrentnode .
8	U NTIL *h*(best_node) < *h*(current_node)
9	I F *h*(current_node)< T_c THEN
10	c urrent_node ← randomlyr e-generatei nitialg roupa ssignment
11	E NDIF
12	UNTIL *h*(current_node) ≥T_c

Table 8 - Algorithm for SVM subset selection

1	LetM ={ SVM_1,S VM_2, …,S VM_N} bet hes eto fc andidateS VMs
2	LetS $core_{max}$ =V (S,M)a ndS et_{max}= M
3	FORi = N -1t o1
4	V $_{max}$= m ax{V(S,M –{S VM_r}) |S VM_r ∈M ,1 ≤r ≤N }
5	I FV (S,M –{ SVM_j})= =V $_{max}$(1 ≤ j ≤N) THEN
6	M = M –{ SVM_j}
7	E NDIF
8	I FV $_{max}$ ≥S $core_{max}$ THEN
9	S $core_{max}$ =V $_{max}$
10	S et_{max}= M
11	E NDIF
12	ENDF OR

Table 9 - The encoding methods of input vectors in the fifteen selected SVMs.

No.	Encoding methodso fi nputve ctors
1	1-gramw ith2pa rtitionedpa rts
2	1-gramw ith3pa rtitionedpa rts
3	1-gramw ith4pa rtitionedpa rts
4	1-gramw ith4pa rtitionedpa rts(applyf eatures electiont oN o.3)
5	1-gramw ith6pa rtitionedpa rts
6	2-gramw ithouta ny gaps
7	2-gramw ithouta ny gaps (apply features electiont oN o.6)
8	2-gramw ithone gap
9	3-gramw ith6m erged groups
10	3-gramw ith7m erged groups
11	3-gramw ith8m erged groups
12	4-gramw ith4m erged groups
13	4-gramw ith4m erged groups
14	4-gramw ith4m erged groups
15	4-gramw ith4m erged groups(apply features electiont oN o.14)