BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS **252,** 63–68 (1998)
ARTICLE NO. RC989498

Using Discriminant Function for Prediction of Subcellular Location of Prokaryotic Proteins

Kuo-Chen Chou and David W. Elrod
Computer-Aided Drug Discovery, Pharmacia & Upjohn, Kalamazoo, Michigan 49007-4940

E-mail: kuo-chen.chou@am.pnu.com

Received September 14, 1998

The discriminant function algorithm was introduced to predict the subcellular location of proteins in prokaryotic organisms from their amino-acid composition. The rate of correct prediction for the three possible subcellular locations of prokaryotic proteins studied by Reinhardt and Hubbard (*Nucleic Acid Research,* 1998, 26:2230–2236) was 90% by the self-consistency test, and 87% by the jackknife test. These rates are considerably higher than the results recently reported by them using the neural network method. Furthermore, the test procedure adopted here is also more rigorous. The core of the current algorithm is the covariance matrix, through which the collective interactions among different amino-acid components of a protein can be reflected. It is anticipated that, owing to the intimate correlation of the function of a protein with its subcellular location, the current algorithm will become a useful tool for the systematic analysis of genome data. © 1998 Academic Press

Key Words: **organelles; amino-acid composition; self-consistency; jackknife; collective interaction.**

The rapidly increasing number of sequences entering into the genome databank has created the need for fully automated methods to analyze them [1]. Knowing the cellular location of a protein is a key step towards understanding its function. Even if the basic function of a protein is known, knowing its cellular location may provide insights as to which pathway an enzyme is involved. The pioneer study by Nakashima and Nishikawa [2] indicated that intra- and extracellular proteins differ significantly in their amino-acid composition. Subsequently two automatic methods for assignment of the subcellular location of proteins according to their amino-acid composition were proposed. One of these [3] is based on Mahalanobis distance [4] which, however, is valid only when the subset sizes in the training dataset are the same or approximately the same [5]; while the other is based on the neural network technique [6] for which it is difficult to give a physical explanation although the results are often successful in practice. For example, as pointed out by King [7], the neural networks methods have “very poor explanatory power” and “they are statistically rather poorly characterized”. Nevertheless, in comparison with [3], the dataset constructed by Reinhardt and Hubbard in [6] is one step forward as reflected by the following features: (a) intracellular proteins are distinguished as cytoplasmic or mitochondrial and eukaryotic and prokaryotic sequences handled separately; (b) all transmembrane proteins are excluded because reliable prediction methods for this group already exist [8]; (c) the number of proteins in each subset (subcellular location) is considerably different as reflecting the reality in cells. In view of this, the Reinhardt and Hubbard dataset can be used to examine the effectiveness of a new prediction algorithm.

DISCRIMINANT FUNCTION

Suppose there are N proteins forming a set S, which is the union of m subsets S_ξ ($\xi = 1, 2, \ldots, m$) each representing a subcellular location. The size of each subset is given by N_ξ ($\xi = 1, 2, 3, \ldots, m$), where N_ξ represents the number of proteins in the subcellular location ξ. Obviously, $N = \sum_{\xi=1}^{m} N_\xi$. The prediction algorithm is based on the correlation between the subcellular location of a protein and its amino-acid composition. Any protein corresponds to a vector or a point in the 20-D (dimensional) space; i.e., it can be described by [9]

$$\mathbf{X}_k^\xi = \begin{bmatrix} x_{k,1}^\xi \\ x_{k,2}^\xi \\ \vdots \\ x_{k,20}^\xi \end{bmatrix}, \qquad (k = 1, 2, \ldots, N_\xi; \quad \xi = 1, 2, 3, \ldots, m) \quad [1]$$

where $x_{k,1}^\xi, x_{k,2}^\xi, \ldots, x_{k,20}^\xi$ are the normalized occurrence-frequencies of the 20 amino acids in the kth

TABLE 1

List of the 997 Prokaryotic Protein Sequences Classified in Three Subcellular Locations as Studied by Reinhardt and Hubbard [6]

(1) 688 Cytoplasmic prokaryotic proteins

SYFA_ECOLI	EFTU_BURCE	PTHP_STAAU	SERC_ECOLI	E4PD_ECOLI	G6PA_BACST	LEU3_LACLA	DLD1_PSEPU	SYG_MYCGE	CFA_ECOLI
SYC_BACSU	TRM9_ECOLI	DHA_BACSU	NODB_RHIME	G3P2_SYNY3	EFTU_CHLTR	DHSS_ANACY	LPLA_ECOLI	KKA4_BACCI	G3P_BUCAP
SYI_PSEFL	UVRC_PSEFL	CH60_ECOLI	DLDH_PSEFL	EFG_AQUPY	PRIS_DESVH	CH10_STAAU	SYIP_STAAU	ARGJ_NEIGO	P115_MYCHR
EFG_SYNY3	LYSR_ECOLI	PHBC_ALCEU	PGK_BACST	EFTU_FLAFE	CLCR_PSEPU	HPRT_RHOCA	PT1_ECOLI	XYLA_BACSU	SYK_CAMJE
PIMT_ECOLI	KAD_MYCGE	SYH_HAEIN	PTFA_BACSU	COBO_PSEDE	NRDG_HAEIN	ASRC_SALTY	CHEY_ECOLI	SOXS_ECOLI	TYSY_HAEIN
XYLA_STRAL	GLPD_ECOLI	CHEA_SALTY	GLYA_CAMJE	SYE_RHIME	SYM_MYCGE	ALKH_ECOLI	PHHC_PSEAE	BTUR_SALTY	KAD_LACLA
SYS_BACSU	SYC_ECOLI	RNB_HAEIN	SYM_BACSU	XYLA_ECOLI	SYI_STAAU	EFG_THEMA	LEU3_ECOLI	DLTA_LACCA	SYW_HAEIN
GLNA_AZOVI	SAOX_BACSP	PRIS_DESDE	OTCC_PSEAE	DLDH_BACSU	CHEW_ECOLI	LON_BACSU	PEPE_ECOLI	RUBR_DESGI	CYAA_BRELI
MALQ_STRPN	EFTU_SPIPL	SYN_MYCGE	PT1_BACST	AMY3_DICTH	DLDH_AZOVI	SYK_BACSU	CYSE_HAEIN	DHSS_SYNP1	CH60_STAAU
GT_ECOLI	GLNA_METCA	SAOX_BACSN	PTCB_ECOLI	APT_MYCGE	CYSB_ECOLI	FABB_HAEIN	EXOA_STRPN	SYS_ECOLI	IPYR_THETH
ACEK_ECOLI	CSCA_ECOLI	EXOA_BACSU	HEMN_HAEIN	G3P_MYCLE	IF1_MYCGE	SYE_HAEIN	SYL_BACSU	SYQ_HAEIN	EFTU_NEIGO
ISP1_BACSU	SYT2_BACSU	BARS_BACAM	NDK_BACSU	IPYR_THEP3	RURE_ACICA	SYY_BACCA	SYP_ECOLI	HPRT_VIBHA	CILB_KLEPN
EFTU_STROR	HMC6_DESVH	DEOC_MYCPI	G3P3_ANAVA	SYGB_HAEIN	SYY_ECOLI	TYRB_ECOLI	AMY2_SALTY	PTMA_STRMU	GCVA_ECOLI
METB_MYCLE	TRM1_ECOLI	GLNA_THIFE	SYK_MYCGE	MELA_SHECO	OTCA_PSEAE	CHEY_BACSU	LON_ECOLI	NODB_RHILP	XYLA_HAEIN
GLYA_ACTAC	SYM_THETH	ADI_ECOLI	G3P_CORGL	MOXR_PARDE	SYQ_ECOLI	SYK2_ECOLI	DLD2_PSEPU	XYLA_LACBR	CILA_HAEIN
DEOC_MYCHO	TYRA_ERWHE	BODG_PSESK	METC_ECOLI	RNE_HAEIN	LEU3_BACSU	SYFA_HAEIN	CAPB_PSEFR	SYD_ECOLI	GLYA_HAEIN
GLNA_ANASP	GLNA_AZOBR	PT1_STRMU	PGK_MYCGE	SELB_ECOLI	35KD_MYCTU	GLNA_THEMA	KAD_YEREN	SYS_HAEIN	ASPG_BACSU
EFTU_CYTLY	PEPT_SALTY	MURF_ECOLI	SYR_MYCGE	NDK_HAEIN	IF2_BACSU	GLNA_LACDE	ALKH_ZYMMO	XYLA_THEET	PHEA_HAEIN
SYS_COXBU	LPLA_MYCGE	RND_ECOLI	METB_HAEIN	EFTS_MYCGE	LEU3_BACCA	TYSY_LACLA	NODA_RHIME	CSPB_BACSU	PTRA_KLEPN
CHEA_ECOLI	METR_ECOLI	DEOC_MYCPN	RIML_ECOLI	NADA_SALTY	CYNR_ECOLI	SYFB_BACSU	PTHP_ALCEU	G3P_MYCGE	PGK_THEMA
PTKB_ECOLI	SYD_THETH	PTNA_ECOLI	G6PI_HAEIN	GSHR_ECOLI	RHAS_ECOLI	HPRT_ECOLI	GSHR_PSEAE	PTFB_BACSU	DHA_BACST
GLN1_RHILV	OTCC_NEIGO	IPYR_MYCGE	PNP_PHOLU	VRPR_SALDU	ACEA_CORGL	GLYA_HYPME	KAD_MYCCA	GLNA_SYNP2	SYM_BACST
EFTU_BACFR	DSVC_DESVH	METC_SALTY	LEU3_THEAQ	ASG1_ECOLI	SYL_ECOLI	ACEA_ECOLI	PHBB_ALCEU	PCP_BACAM	SYFB_HAEIN
G3P_HAEIN	SYY2_BACSU	CHEZ_ECOLI	PTGA_ECOLI	SYE_THETH	KAD_MICLU	KAD_ECOLI	GLNA_BACSU	TAGF_BACSU	RIMI_HAEIN
LEU3_LEPIN	CILA_KLEPN	XYLA_CLOTS	CHEW_BACSU	XYLA_THENE	GLN2_BRAJA	SYN_HAEIN	AFQ1_STRCO	SYK_THETH	HEMN_ECOLI
SYR_MYCLE	PTHP_MYCCA	SYR_ECOLI	GLNA_CLOAB	EFTU_CORGL	SYT_MYCGE	VIRG_AGRRA	SYP_MYCGE	SYV_LACCA	SYH_MYCGE
ACEA_MYCLE	XYLA_THETH	SYN_ECOLI	EFTU_MYCGA	EFTU_BACSU	XYLA_STRRO	PTLA_LACLA	LON_BACBR	KAD_BACST	TRM2_ECOLI
SYW_ECOLI	EFTU_TAXOC	SCRB_VIBAL	THIL_ZOORA	ENP2_BACSH	METX_HAEIN	SYFA_BACSU	EFTU_THEAQ	FES_ECOLI	GT_PROMI
PHOH_ECOLI	SYW_BACST	EFG_MICLU	PNP_ECOLI	SYM_HAEIN	IADA_ECOLI	PTHP_STRMU	G6PI_ZYMMO	RURE_PSEOL	AMPL_RICPR
XGPT_ECOLI	DEOC_ECOLI	LEU3_BACME	EFTU_FLESI	PT1_HAEIN	GLN1_STRVR	HEM6_ECOLI	PHEA_PSEST	PT1_BACSU	SYD_HAEIN
SYI_MYCGE	HOXU_ALCEU	XYLA_LACPE	IF2_ECOLI	NODB_BRASP	ENO_ECOLI	SYL_HAEIN	DHA_BACSH	VGB_STAAU	PAPB_ECOLI
CHMU_BACSU	PGK_MYCLE	USPA_ECOLI	AMPR_CITFR	HEMN_RHOSH	DLDH_ECOLI	SCRB_KLEPN	SYP_CHLTR	NAHR_PSEPU	EFTU_MYCHO
XYLA_ARTS7	CILB_HAEIN	DEOC_BACSU	MASY_ECOLI	THIL_THIVI	NFRC_ECOLI	GLNA_ECOLI	G6PB_BACST	THIL_BACSU	G3P_ZYMMO
CSPA_ECOLI	O16G_BACTR	ACKA_MYCGE	SYT_ECOLI	DDLB_ECOLI	LEU3_BACCO	IF2_THETH	PROB_BACSU	PCP_STRPY	TYRA_HAEIN
SCRB_SALTY	THGA_ECOLI	MALZ_ECOLI	GLYA_ECOLI	PTWX_ECOLI	GLYA_BACSU	PHOB_ECOLI	PTMA_STACA	PT1A_ECOLI	TYSY_LACCA
SYW_BACSU	IF2_HAEIN	GLNA_PROVU	G3P1_SYNY3	DDLA_ECOLI	XYS3_PSEPU	G6PI_MYCGE	RNS_ECOLI	ASRA_SALTY	TRXB_ECOLI
PCP_BACSU	SYFA_MYCGE	PTHP_ECOLI	ASRB_SALTY	LEU3_CLOPA	GLN1_FRAAL	RNC_HAEIN	RIMJ_ECOLI	PTHP_BACSU	IF1_LACLA
AMPR_ENTCL	EFG_MYCGE	PROB_ECOLI	NDK_ECOLI	AAT_HAEIN	LEU3_AGRTU	AMY2_ECOLI	NIRD_ECOLI	APT_PSEAE	DLD3_PSEPU
AMPD_CITFR	SPAR_BACSU	SYS_MYCGE	OTC2_BACSU	PTRB_KLEPN	LON_MYCGE	LEU3_BUCAP	GT_HAEIN	XYLA_STAXY	SYE_BACSU
HPRT_LACLA	UREE_HELPY	OTC_HAEIN	DCP_ECOLI	GSHR_HAEIN	DLDH_MYCGE	SYV_HAEIN	GLNA_VIBAL	SYM_ECOLI	SYK1_ECOLI
PT1_STRSL	UVRC_MYCGE	SYY_MYCGE	CH10_STAEP	G3P_BACME	INVA_ZYMMO	SYV_ECOLI	IF2_BACST	AMPR_YEREN	GLNA_FREDI
EFTU_RICPR	PTHA_ECOLI	CAFA_ECOLI	XYS2_PSEPU	PHOB_PSEAE	SYL_MYCGE	FABB_ECOLI	PT1_STACA	OTCA_PSESH	PTFA_HAEIN
DEOC_HAEIN	UVRB_MICLU	ACEA_RHOFA	APT_HAEIN	G3P2_ANAVA	RNB_ECOLI	PTFA_ECOLI	EFTU_WOLSU	EFT2_STRRA	AGRA_STAAU
PGK_ECOLI	EFG_SPIPL	UREE_KLEAE	UBIC_ECOLI	CAIB_ECOLI	PGK_HAEIN	CILG_HAEIN	NRDG_ECOLI	CITR_BACSU	SYD_MYCGE
SYGB_ECOLI	NODA_RHILV	THIK_ECOLI	UVRC_ECOLI	DEOC_MYCGE	PTGA_MYCCA	LON2_MYXXA	SYA_MYCGE	GLN2_RHILP	G6PI_ECOLI
SYH_MYCLE	PHEA_ERWHE	IF2_MYCGE	CYPC_ECOLI	CYSR_SYNP7	GLNA_BACCE	PROB_CORGL	RHAR_ECOLI	PCP_PSEFL	PECS_ERWCH
EFTS_THETH	SYV_MYCGE	DCP_SALTY	ALKK_PSEOL	EFTU_MYCLE	SYFB_THETH	SYK_HAEIN	CRL_ECOLI	EFTS_HAEIN	O16G_BACCE
CYPB_HAEIN	NODB_RHILV	SYR_HAEIN	PGK_CORGL	TREC_ECOLI	EFTS_SPICI	LON_HAEIN	SYW_MYCGE	SYH_STREQ	TETX_BACFR
PT1_MYCGE	LPSZ_RHIME	SYFA_THETH	SYE_BACST	PTMA_ENTFA	GLNA_HAEIN	EFTU_MYCGE	EFTU_SHEPU	TCPN_VIBCH	AMY2_DICTH
TAGE_BACSU	SYT1_BACSU	SYP_HAEIN	EFTS_ECOLI	DEXB_STRMU	EFTU_BRELN	SYY1_BACSU	SYI_CAMJE	G3P_THEMA	PHBB_CHRVI
AMPD_ECOLI	HOXH_ALCEU	RHAS_SALTY	CAFA_HAEIN	PT1_ALCEU	KAD_BORPE	EFTU_DEISP	SYY_THIFE	KDSA_ECOLI	SYK_MYCHO
SAOX_CORSP	CHEW_ENTAE	METK_ECOLI	FUMC_BRAJA	PTHP_BACST	RND_HAEIN	HOXF_ALCEU	GLN2_RHIME	KAD_BACSU	ILVY_ECOLI
HPRT_HAEIN	NHAR_ECOLI	IPYR_HAEIN	SYC_HAEIN	SYE_AZOBR	BTUR_ECOLI	FDHD_WOLSU	PROB_HAEIN	OMPR_ECOLI	RUBR_DESVH
VIRG_AGRT6	CILG_KLEPN	BGLB_MICBI	HEM6_PSEAE	LEU3_SPIPL	GLYA_MYCGE	PRSX_ECOLI	METC_HAEIN	PFLA_ECOLI	PTLA_STAAU
SYFB_MYCGE	EFG_THETH	PTLA_STRMU	ACKA_ECOLI	SYV_BACSU	NODA_BRASP	SYT_HAEIN	NODA_AZOCA	IF1_BACSU	OTCA_MYCBO
SLYD_ECOLI	GLNA_NEIGO	PT1_MYCCA	PGK_BACME	G3P_THEAQ	EFG_HAEIN	GSHR_BURCE	XYLA_ACTMI	CYSE_ECOLI	SERC_HAEIN
GSHR_ANASP	G3P1_ANAVA	EFT1_STRCO	TYSY_MYCTU	CHEB_ECOLI	EFTU_CHLVI	SYH_ECOLI	ACKA_HAEIN	PHEA_ECOLI	RIMI_ECOLI
UGPQ_ECOLI	DAPD_ECOLI	EFTU_ECOLI	CATR_PSEPU	CH10_ECOLI	KAD_HAEIN	OTC1_ECOLI	DTXR_CORDI	XYLS_PSEPU	TAGD_BACSU
EFG_ECOLI	GLN2_FRAAL	CYPB_ECOLI	XYLA_KLEAE	RF3_ECOLI	SYT_BUCAP	HPRT_MYCGE	AAT_ECOLI	SYC_MYCGE	METX_MYCGE
CSPB_BACCL	ISPA_BACST	G3P2_RHOSH	TYRA_ECOLI	IPYR_ECOLI	PTHP_ENTFA	FRZC_MYXXA	SYA_ECOLI	IPI_BACSU	PEPX_LACLA
LEU3_HAEIN	RNC_ECOLI	RF3_HAEIN	PTLA_LACCA	ARAC_CITFR	G3P1_ECOLI	KAD_PARDE	TYSY_ECOLI	THIL_CLOAB	VDH_STRCO
DAPD_ACTPL	G3P_BACCO	SYI_ECOLI	MASY_CORGL	SYR_BRELA	AMPR_PSEAE	IF1_MYCBO	PTKA_ECOLI	SYD_MYCLE	METB_ECOLI
METC_BORAV	PTH_ECOLI	PTCA_ECOLI	TFDR_ALCEU	SYS_THETH	PROB_SERMA	IF1_ECOLI	IF2_ENTFC	RNC_MYCGE	SAOX_ARTSP
TYSY_MYCGE	THIL_CHRVI	SAOX_STRSQ	HLYX_ACTPL	ISPA_HAEIN	SYV_BACST	EFTU_ANANI	CATA_PROMI	GAL_PSEFL	PHBB_ZOORA
TREC_BACSU	SYFB_ECOLI	XGPT_HAEIN	EFG_MYCLE	AMPR_RHOCA	PEPE_SALTY	EFT3_STRCO	NODA_RHILT	DLDH_HAEIN	SYGA_ECOLI
SYE_MYCGE	ASPG_BACLI	SELB_HAEIN	NODB_AZOCA	PILB_PSEAE	PFLB_ECOLI	ENO_ZYMMO	SYA_HAEIN	EFTU_HERAU	UVRC_BACSU
AMY1_DICTH	RNE_ECOLI	EFG_ANANI	O16G_BACSP	EFTU_MICLU	HPRT_BACSU	PTWB_ECOLI	SYE_ECOLI	TFDT_ALCEU	FOSB_STAEP
MLER_LACLA	PMBA_ECOLI	LON1_MYXXA	THIL_ALCEU	HOXY_ALCEU	APT_ECOLI	APT_PSEST	NODB_RHILT	UVRB_ECOLI	ARAC_ERWCH
MALQ_ECOLI	PAPX_ECOLI	OTC2_ECOLI	AACA_STAAU	PGK_THETH	NEUA_ECOLI	DLDH_BACST	GLPD_BACSU		

(2) 107 Extracellular prokaryotic proteins

SPI_BACBR	PRTB_ERWCH	THER_BACST	PRSG_ECOLI	GTFB_STRMU	GTF1_STRDO	CHOD_BREST	HRPZ_PSESY	NPRV_VIBPR	EMPA_VIBAN
XYNA_STRLI	NPRE_BACCL	PRT1_ERWCA	DEXT_ARTSP	GTFC_STRMU	NPRE_BACBR	PAPA_ECOLI	SNPA_STRSQ	RN_BACCO	PHL_LEPIN
EBA3_FLAME	XYNB_STRLI	PELE_ERWCH	BPRX_BACNO	PRTC_ERWCH	HLT_VIBPA	PELD_ERWCH	API_ACHLY	SEPA_STAEP	GTFD_STRMU
PROB_STRAG	PELF_ERWCH	AMYB_BACPO	SNPA_STRCO	PIL5_ECOLI	PHL1_BACCE	APRA_PSEAE	LKTA_PASHA	THET_THEVU	A85B_MYCAV
XYNC_PSEFL	HRPN_ERWAM	XYNC_STRLI	STRK_STRGR	PELB_ERWCA	SNPA_STRLI	PROA_XANCP	PRTA_ERWCH	NUC_STAHY	PIL1_SALTY
PRTS_SERMA	PELA_ERWCH	PELA_ERWCA	PEL3_ERWCA	SMP_SERMA	PHL3_BACCE	NPRE_BACSU	AMT4_PSESA	CHOD_STRSQ	PRTG_ERWCH
NUC_SERMA	PEL_BACSU	PAPH_ECOLI	LIPE_AERHY	EBA1_FLAME	SUBF_BACSU	AMT4_PSEST	SACB_STRMU	DRNE_VIBCH	NUCB_BACSU
PAPG_ECOLI	PELB_ERWCH	PELC_ERWCH	NPRE_BACAM	NPRM_BACME	PIL1_ECOLI	PBPA_STRPN	LSTP_STAST	AMT6_BACS7	RNBR_BACAM
SUBV_BACSU	LIP_PSESP	PRSE_ECOLI	PHB_ALCFA	PEL1_ERWCA	A85B_MYCKA	AMYR_BACS8	NPRS_BACST	NPRE_BACCE	SODF_MYCTU
SUBE_BACSU	AGAR_STRCO	COMX_BACSU	ELAS_PSEAE	PAPF_ECOLI	LSTP_STASI	TCPA_VIBCH	BPRV_BACNO	A85B_MYCLE	EBA2_FLAME
A85B_MYCBO	PAPE_ECOLI	MPR_BACSU	DRNE_AERHY	PRTT_SERMA	CYAA_BORPE	PROA_LEGPN			

TABLE 1—*Continued*

(3) 202 Periplasmic prokaryotic proteins

AGP_ECOLI	AZUR_PSEPU	FANE_ECOLI	PHEC_PSEAE	AMO_ECOLI	AZUP_METEX	TRAF_ECOLI	PHNS_DESVM	SODC_CAUCR	TBPA_ECOLI
MALM_ECOLI	PPA_ZYMMO	SPEA_ECOLI	PPB_SERMA	DHML_METEX	DHMH_PARDE	PHON_PROST	C553_BRAJA	DHET_ACEAC	DHM1_PARDE
PHFL_DESVH	TORA_ECOLI	NIR_PSESP	PRC_ECOLI	BLAC_RHOCA	OSMY_ECOLI	SUBI_SYNY3	PHON_SALTY	AZUR_ALCFA	SFUA_SERMA
FLGI_CAUCR	NIR_PSEAR	TRAW_ECOLI	HFB1_HAEIN	C552_BRAJA	FRDA_SHEPU	POTF_ECOLI	TRBB_ECOLI	FBP_HAEIN	AMY_THETU
PHNL_DESVM	PHNL_DESFR	MDOG_ECOLI	GALM_ACICA	CLPE_ECOLI	PHOC_MORMO	PBP7_ECOLI	KSD1_ECOLI	MYFB_YEREN	DCTP_RHOCA
COPC_PSESM	C552_PSEST	AZUR_ALCDE	AZUR_ALCSP	FECR_ECOLI	HELX_RHOCA	PHFS_DESVO	COPA_PSESM	AZUR_BORBR	PHSL_DESBA
NRFA_ECOLI	BRAC_PSEAE	NOSZ_PSEST	PAPJ_ECOLI	LOLA_ECOLI	HISJ_SALTY	TRH1_ECOLI	AZUR_PSEAE	C553_PARDE	ASG2_ECOLI
TESA_ECOLI	GLPQ_ECOLI	YTFQ_ECOLI	THTR_SYNP7	NANH_CLOPE	INH_PSEAE	C551_PSEST	C551_PSEAE	TCPG_VIBCH	DSBC_ECOLI
FLAA_SPIAU	NANH_CLOSE	TOLB_ECOLI	DHM2_PARDE	AZUR_PSEFD	C550_PSEST	GGT_ECOLI	OPPA_SALTY	AMO_KLEAE	PHNS_DESFR
DHM1_METEX	CYPH_ECOLI	CHVE_AGRTU	PHNS_DESGI	GUNB_PSEFL	FBP_NEIGO	RBSB_ECOLI	NANH_CLOSO	PHF1_CLOPA	AZUP_ALCFA
NIR_ALCFA	PAC_ECOLI	MODA_ECOLI	TREA_ECOLI	RHIC_RHILV	DHMH_THIVE	CGKA_ALTCA	NIRS_PSEAE	PROX_ECOLI	INH_ERWCH
LIVJ_CITFR	PHFS_DESVH	AZUR_PSEDE	OCCT_AGRT6	HTRA_ECOLI	ECPD_ECOLI	SUBI_SYNP7	TRAU_ECOLI	AMY1_ECOLI	SUFI_ECOLI
ALBR_KLEOX	PHSS_DESBA	UGPB_ECOLI	MEPA_ECOLI	C553_DESVM	LACE_AGRRD	CHMU_ERWHE	MALE_ECOLI	ARAF_ECOLI	FIMC_ECOLI
GLNH_ECOLI	DPPA_ECOLI	NIR_ACHCY	NOSD_PSEST	DGAL_CITFR	FEPB_ECOLI	OPPA_ECOLI	MODB_AZOVI	DHML_PARDE	PSTS_ECOLI
TRBC_ECOLI	AZU2_METJ	ICSB_SHIFL	CYSD_CHRVI	POTD_ECOLI	TBPA_HAEIN	PPA_ECOLI	LIVK_ECOLI	FLA1_BORBU	PICP_PSESP
FLA1_TREHY	PPB4_BACSU	ALGL_PSEAE	AZUR_PSEFB	CYSP_ECOLI	DHM1_METME	FER2_DESDN	AZUR_PSEFC	PHFL_DESVO	SODC_BRUAB
XYLF_ECOLI	PTR_ECOLI	C553_DESVH	DSBE_ECOLI	SODC_PHOLE	FLB2_TREHY	AZUP_ACHCY	RUS1_THIFE	PELP_ERWCA	E13B_OERXA
C553_DESDN	AZU1_METJ	NAPA_ALCEU	FLGI_SALTY	FECB_ECOLI	BGLX_ECOLI	DSBA_HAEIN	CN16_ECOLI	PRC_HAEIN	C562_ECOLI
PPB3_BACSU	USHA_ECOLI	DSBC_ERWCH	ECOT_ECOLI	DHGA_ACICA	MRKB_KLEPN	HEP1_FLAHE	NAPB_ALCEU	DHM2_METEX	PPB_ECOLI
PHNL_DESGI	NOSZ_PSEAE	NIRS_PSEST	CAFM_YERPE	SUBI_ECOLI	NUCM_ERWCH	DSBA_ECOLI	PAPD_ECOLI	PELP_YERPS	PPCE_FLAME
ARGT_SALTY	DHSU_CHRVI								

Note. The codes are according to the SWISS-PROT Data Bank.

protein $\mathbf{X}_k^\xi$ of the ξth subcellular location. The *standard vector* for the subcellular location ξ is defined by

$$\mathbf{X}^\xi = \begin{bmatrix} x_1^\xi \\ x_2^\xi \\ \cdot \\ \cdot \\ \cdot \\ x_{20}^\xi \end{bmatrix}, \quad (\xi = 1, 2, 3, \ldots, m) \qquad [2]$$

where

$$x_i^\xi = \frac{1}{N_\xi} \sum_{k=1}^{N_\xi} x_{k,i}, \quad (i = 1, 2, \ldots, 20). \qquad [3]$$

Suppose $\mathbf{X}$ is a protein whose subcellular location is to be predicted. It also corresponds to a point $(x_1, x_2, \ldots, x_{20})$ in the 20-D space, where x_i has the same meaning as $x_{k,i}^\xi$ but is associated with protein $\mathbf{X}$ instead of $\mathbf{X}_k^\xi$. Thus, the current algorithm can be formulated as follows.

The similarity between the standard vector $\mathbf{X}^\xi$ and the protein $\mathbf{X}$ is characterized by the Bayes discriminant function, as defined by [10]

$$F(\mathbf{X}, \mathbf{X}^\xi) = D^2(\mathbf{X}, \mathbf{X}^\xi) + \ln(\lambda_2^\xi \lambda_3^\xi \lambda_4^\xi, \ldots, \lambda_{20}^\xi). \qquad [4]$$

The first term is the squared Mahalanobis distance between $\mathbf{X}^\xi$ and $\mathbf{X}$ [4, 11]:

$$D^2(\mathbf{X}, \mathbf{X}^\xi) = (\mathbf{X} - \mathbf{X}^\xi)^{\mathbf{T}} \mathbf{C}_\xi^{-1} (\mathbf{X} - \mathbf{X}^\xi), \quad (\xi = 1, 2, 3, \ldots, m) \qquad [5]$$

where $\mathbf{C}_\xi$ is the covariance matrix for subset S^ξ, given by

$$\mathbf{C}_\xi = \begin{bmatrix} c_{1,1}^\xi & c_{1,2}^\xi & \cdots & c_{1,20}^\xi \\ c_{2,1}^\xi & c_{2,2}^\xi & \cdots & c_{2,20}^\xi \\ \vdots & \vdots & \ddots & \vdots \\ c_{20,1}^\xi & c_{20,2}^\xi & \cdots & c_{20,20}^\xi \end{bmatrix}, \qquad [6]$$

and the superscript $\mathbf{T}$ is the transposition operator; $\mathbf{C}_\xi^{-1}$ is the inverse matrix of $\mathbf{C}_\xi$. The matrix elements $c_{i,j}^\xi$ in eq.6 are given by

$$c_{i,j}^\xi = \frac{1}{N_\xi - 1} \sum_{k=1}^{N_\xi} [x_{k,i}^\xi - x_i^\xi][x_{k,j}^\xi - x_j^\xi], \quad (i, j = 1, 2, \ldots, 20). \qquad [7]$$

Note that, different from the covariant matrices formulated in [4], a denominator $N_\xi - 1$ is incorporated in the above equation. The second term of eq.4 reflects the difference of covariance matrices for different subcellular locations, in which λ_i^ξ is the ith eigenvalue of the covariance matrix $\mathbf{C}_\xi$ $(i = 2, 3, 4, \ldots, 20)$. It can be proved that, for the covariance matrix $\mathbf{C}_\xi$ as defined by eq.7, there are no negative eigenvalues. It can also be proven [9] that $\mathbf{C}_\xi$ has one, and only one, eigenvalue (represented by λ_1^ξ) equal to zero; i.e., $\lambda_1^\xi = 0$. Incorporation of the term $\ln(\lambda_2^\xi \lambda_3^\xi \lambda_4^\xi, \ldots, \lambda_{20}^\xi)$ into the discriminant function, together with the denominator $N_\xi - 1$ into the covariant matrices, is very important, especially when the subset sizes in the training dataset are much different [5]. It is because of the second term that the discriminant function F as defined by eq.4 is no longer a distance because it does not satisfy the condition of $F(\mathbf{X}, \mathbf{X}^\xi) = 0$ when $\mathbf{X} \equiv \mathbf{X}^\xi$, and also it may have a negative value, obviously in conflict with the classical definition that a distance must satisfy positivity, symmetry, and the triangular inequality.

TABLE 2

Predicted Results for the Three Possible Subcellular Locations of the 997 Prokaryotic Proteins in Table 1

	Rate of correct prediction for each subcellular location			
Test method	1. Cytoplasmic[a]	2. Extracellular[a]	3. Periplasmic[a]	Overall rate of correct prediction
Self-consistency	$\frac{643}{688} = 93.5\%$	$\frac{94}{107} = 87.9\%$	$\frac{164}{202} = 81.2\%$	$\frac{901}{997} = 90.4\%$
Jackknife	$\frac{630}{688} = 91.6\%$	$\frac{86}{107} = 80.4\%$	$\frac{146}{202} = 72.3\%$	$\frac{862}{997} = 86.5\%$

[a] The number of proteins in this group has one or two proteins more than that of Table 1 of ref.6. This is because during the training process performed by Reinhardt and Hubbard all groups had to have a number of sequences dividable by three. As a consequence they left out 1 or two at the end of those groups if the number of proteins therein cannot be perfectly divided by three (personal communication with Dr. Reinhardt).

Thus, the prediction rule is formulated by

$$F(\mathbf{X}, \mathbf{X}^{\chi}) = \mathbf{Min}\{F(\mathbf{X}, \mathbf{X}^{1}), F(\mathbf{X}, \mathbf{X}^{2}), F(\mathbf{X}, \mathbf{X}^{3}), \ldots, F(\mathbf{X}, \mathbf{X}^{m})\} \quad [8]$$

where χ can be 1, 2, 3, . . . , or m, and the operator **Min** means taking the least one among those in the parentheses, then the superscript χ of eq.8 is the predicted cellular location for the protein **X**. If there is a tie case, ξ is not uniquely determined, but that did not occur in our dataset.

RESULTS AND DISCUSSION

To show the power of the current prediction algorithm, the comparison was made with the best result reported by the previous investigators. According to a recent report by Reinhardt and Hubbard [6], for the 997 prokaryotic proteins classified in three different subcellular locations (Table 1), the rate of correct prediction by the neural network method was 81%. This is the highest accuracy rate so far reported about the prediction of protein cellular location. Now for the same dataset, we used the discriminant function algorithm to perform prediction. The prediction quality

TABLE 3

The Standard Vector and Eigenvalue Set Derived from the Dataset of Table 1 for Each of the Three Subcellular Locations of Prokaryotic Proteins

Standard vector				Eigenvalue set			
Amino acid code	1. Cytoplasmic $\mathbf{X}^1$	2. Extracellular $\mathbf{X}^2$	3. Periplasmic $\mathbf{X}^3$	Order i	1. Cytoplasmic $\lambda_i^1 \times 10^5$	2. Extracellular $\lambda_i^2 \times 10^5$	3. Periplasmic $\lambda_i^3 \times 10^5$
A	0.089	0.098	0.106	1	0	0	0
C	0.010	0.007	0.012	2	0.5	0.7	0.6
D	0.060	0.058	0.061	3	4.2	2.7	4.9
E	0.075	0.037	0.050	4	6.1	4.3	8.7
F	0.039	0.034	0.036	5	8.0	5.2	9.0
G	0.074	0.096	0.081	6	10.1	6.6	12.3
H	0.024	0.017	0.019	7	11.3	8.7	14.2
I	0.063	0.046	0.046	8	13.9	10.4	15.8
K	0.060	0.054	0.070	9	16.1	11.3	16.6
L	0.092	0.070	0.082	10	16.5	15.5	19.1
M	0.026	0.020	0.028	11	21.3	18.8	23.6
N	0.039	0.065	0.046	12	23.2	22.6	26.7
P	0.041	0.038	0.050	13	25.8	29.0	32.0
Q	0.037	0.040	0.041	14	28.5	33.0	38.9
R	0.053	0.037	0.036	15	31.4	38.0	43.0
S	0.050	0.081	0.061	16	38.1	43.5	49.3
T	0.053	0.071	0.059	17	48.3	57.0	67.4
V	0.074	0.068	0.071	18	66.0	80.4	73.1
W	0.010	0.017	0.014	19	99.3	100.6	112.5
Y	0.029	0.044	0.032	20	146.0	148.7	128.3

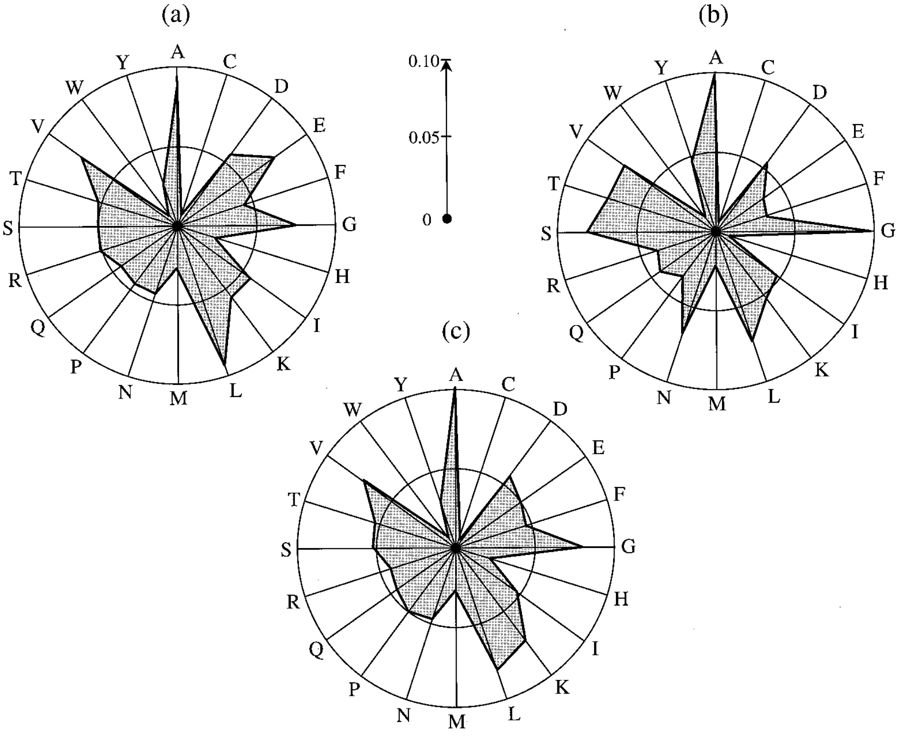

FIG. 1. Radar diagrams to show the difference of the 20-D standard vectors, i.e. the average amino acid compositions which distinguish the subcellular locations of (a) cytoplasmic prokaryotic proteins, (b) extracellular prokaryotic proteins, and (c) peroplasmic prokaryotic proteins. Amino acids are denoted by their single-letter codes (see Table 3).

was examined by the standard testing procedure in statistics [12] that consists of the self-consistency and jackknife tests. In the former, the subcellular location for each protein in a given dataset was predicted using the parameters derived from the same dataset, the so-called training dataset; while in the latter, each protein in the training dataset was singled out in turn as a “test protein” and all the rule-parameters were derived from the remaining proteins. Compared with the independent dataset test and sub-sampling test often adopted in biology, the jackknife test is thought the most effective method for cross-validation in statistics [12]. This is because in the independent dataset test, the selection of a testing dataset is arbitrary, and the accuracy thus obtained lacks an objective criterion unless the testing dataset is sufficiently large [9]. As for the subsampling test in which a given dataset is divided into two or three subsets, the problem is that the number of possible divisions might be too large to be handled. For example, in the treatment by Reinhardt and Hubbard [6], proteins in each group of Table 1 were equally divided into three subgroups. Thus, the number of possible divisions would be $\Psi = \Psi_1 \times \Psi_2 \times \Psi_3$, where $\Psi_1 = \frac{687!}{229!229!229!}$, $\Psi_2 = \frac{105!}{35!35!35!}$, and $\Psi_3 = \frac{201!}{67!67!67!}$. Of Ψ_1, Ψ_2, and Ψ_3, the smallest is $\Psi_2 \simeq 9.8 \times 10^{47}$, indicating the number of possible divisions would be $\Psi \gg 10^{141}$! This is an astronomical figure, which is too large to be handled by any existing computers. Hence in any practical sub-sampling tests as carried out in [6], only a very small fraction of the possible divisions were investigated, and the results thus obtained would certainly bear considerable arbitrariness. Accordingly, the testing procedure adopted here is much more objective and rigorous.

The predicted results by self-consistency and jackknife tests for the 997 proteins of Table 1 are given in Table 2, from which we can see that the overall rate of correct prediction is 90% by self-consistency test, and 87% by jackknife test. Both are considerably higher than the prediction accuracy of 81% obtained by the neural network method as reported in [6]. Likewise, better prediction quality was also obtained by using the current method for all the other datasets constructed for studying cellular location of proteins.

Therefore, from both the rationality of testing procedure and the accuracy of test results, the introduction of the discriminant function algorithm as presented in this paper can significantly improve the prediction quality.

To show the difference in amino acid compositions that distinguish the subcellular locations of proteins, the 20-D standard vector derived from the proteins in Table 1 for each of the three subcellular locations is given in Table 3. Meanwhile, to provide an intuitive picture, each such 20-D standard vectors is projected onto a 2-D radar diagram as given in Fig.1. Furthermore, the 20 eigenvalues for each of the three corresponding covariance matrices are also given in Table 3 that might be of use for investigating the component-coupled effects at a deeper level, especially for understanding the important contribution from the second term of eq.4. This is a vitally important term for dealing with the case where the sizes of subsets are different. However, such an important term as well as the denominator $N_\xi - 1$ in eq.7 were not included in the original least Mahalanobis distance algorithm [4] although good results were still yielded because the case studied there consisted of the same-sized subsets. It is very important to realize this; otherwise, the prediction algorithm might be misused, leading to poor results and an incorrect conclusion.

The essence of the discriminant function algorithm is in the covariance matrix (eq.6), which reflects the collective interactions among different amino-acid components of a protein that actually dictate its final folding state or conformation. On the other hand, different subcellular compartments will provide different optimal environments for some special protein conformations. It is based on such an internal relationship that the current prediction algorithm is established. It is anticipated that with continuously updating the training dataset by incorporating more protein sequences and increasing the accuracy of locational classification, the prediction quality will be further improved. Since the possible function of a protein is restricted by its subcellular location, the powerful prediction algorithm developed here may become a useful vehicle for systematic analysis of the wealth of rapidly increasing data being provided by large scale genome projects.

ACKNOWLEDGMENT

The authors are indebted to Dr. A. Reinhardt for providing the data of Table 1 for testing the discriminant function algorithm.

REFERENCES

1. Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.-C., and Herrmann, R. (1996) *Nucleic Acids Res.* **24,** 4420–4449.
2. Nakashima, H., and Nishikawa, K. (1994) *J. Mol. Biol.* **238,** 54–61.
3. Cedano, J., Aloy, P., Pérez-pons, J. A., and Querol, E. (1997) *J. Mol. Biol.* **266,** 594–600.
4. Chou, K. C. (1995) *Proteins: Structure, Function and Genetics* **21,** 319–344.
5. Chou, K. C., Liu, W., Maggiora, G. M., and Zhang, C. T. (1998) *Proteins: Structure, Function and Genetics* **31,** 97–103.
6. Reinhardt, A., and Hubbard, T. (1998) *Nucleic Acids Res.* **26,** 2230–2236.
7. King, R. D. (1996) In Sternberg, M. J. E. (Ed.), *Protein Structure Prediction: A Practical Approach,* IRL Press, Oxford, pp. 79–97.
8. Rost, B., Casadio, R., Fariselli, P., and Sander, C. (1995) *Protein Science* **4,** 521–533.
9. Chou, K. C., and Zhang, C. T. (1995) *Critical Reviews in Biochemistry and Molecular Biology* **30,** 275–349.
10. Duda, R. O., and Hart, P. E. (1973) *Pattern Classification and Scene Analysis,* Chap.2, John Wiley & Sons, New York.
11. Mahalanobis, P. C. (1936) *Proc. Natl. Inst. Sci. India* **2,** 49–55.
12. Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979) *Multivariate Analysis,* pp. 322, 381, Academic Press, London.