Protein Engineering vol.10 no.6 pp.673–676, 1997

Prediction of transmembrane α-helices in prokaryotic membrane proteins: the dense alignment surface method

Miklos Cserzö[2,3], Erik Wallin[1], Istvan Simon, Gunnar von Heijne[1], Arne Elofsson[1]

Institute of Enzymology, Biological Research Center Hungarian Academy of Sciences, PO Box 7, H-1518 Budapest, Hungary and [1]Department of Biochemistry, Stockholm University, S-106 91 Stockholm, Sweden
[2]Present address: University of Birmingham, School of Biochemistry, Edgbaston, Birmingham B15 2TT, UK

[3]To whom correspondence should be addressed

A new, simple method for predicting transmembrane segments in integral membrane proteins has been developed. It is based on low-stringency dot-plots of the query sequence against a collection of non-homologous membrane proteins using a previously derived scoring matrix [Cserzö *et al.*, 1994, *J. Mol. Biol.*, 243, 388–396]. This so-called dense alignment surface (DAS) method is shown to perform on par with earlier methods that require extra information in the form of multiple sequence alignments or the distribution of positively charged residues outside the transmembrane segments, and thus improves prediction abilities when only single-sequence information is available or for classes of membrane proteins that do not follow the 'positive inside' rule.

Keywords: transmembrane α-helices/prokaryotic membrane proteins

Introduction

Transmembrane helices in integral membrane proteins are composed of stretches of 15–30 predominantly hydrophobic residues separated by polar connecting loops (von Heijne, 1994). A number of algorithms designed to locate putative transmembrane helices in the primary amino acid sequence have been developed, and current methods can identify around 90–95% of all true transmembrane segments with an over-prediction rate of only a few percent (von Heijne, 1992; Jones *et al.*, 1994; Persson and Argos, 1996; Rost *et al.*, 1995; Rost *et al.*, 1996). The best results have been obtained when multiply aligned sequences can be analyzed; however, in many cases there are no homologues in the database and improvements in single-sequence prediction performance are thus important.

Recently, the so-called dense alignment surface (DAS) method was introduced in an attempt to improve sequence alignments in the G-protein coupled receptor family of transmembrane proteins (Cserzö *et al.*, 1994). We have now generalized this method to predict transmembrane segments in any integral membrane protein without the need for multiple-sequence information, and find that it performs on par with the best multiple-alignment based schemes when tested on a set of prokaryotic inner membrane proteins with known topologies.

Materials and methods

Experimental hydrophobicity profiles

A test set of 44 prokaryotic transmembrane proteins with experimentally determined topologies was selected from the SwissProt database (Bairoch and Boeckmann, 1991), Table I. These proteins contain 262 transmembrane segments and a total number of 15 467 residues. For four of the proteins the annotations in the database were found to contain erroneous information compared with the published data. In KDPD_ECOLI (Zimmann *et al.*, 1995) two false segments were indicated (25–45 and 841–861). In TOLQ_ECOLI (Kampfenkel and Braun, 1993; Vianney *et al.*, 1994) the location of the first transmembrane segment was wrong (23–43 instead of 9–36). The following transmembrane segments were missing in the database annotations: GLPT_ECOLI: 28–44, 65–87, 98–113 and 293–310 (Gött and Boos, 1988); SECD_ECOLI: 476–497 and 586–605 (Pogliano and Beckwith, 1994); TOLQ_ECOLI: 127–159 and 162–191 (Kampfenkel and Braun, 1993; Vianney *et al.*, 1994). The number and location of the transmembrane segments were corrected manually for these sequences. One should note that, while the membrane topology—after the proper correction—is reliable, the precise ends of the transmembrane segments are only approximate.

From the corrected feature tables, topology profiles were generated by setting the profile value to 1 for all residues in transmembrane segments and to 0 for the rest of the sequence. We refer to these profiles as 'experimental'.

The redundancy of the test set was checked by pairwise alignment of the sequences ('gap' tool of Wisconsin Package Version 8.1, Genetics Computer Group, Madison, WI; end-weight switch on, standard scoring matrix, default gap penalty). In one case, the percentage identity was 39%, all other pairwise identities were below 30%.

DAS hydrophobicity profiles

The 'dense alignment surface' (DAS) method is based on a traditional dot-plot of two proteins (Cserzö *et al.*, 1994). If two segments of a certain length of the two proteins have a similarity score with a significance higher than a certain cut-off, that region is marked on the dot-plot. DAS uses the RReM scoring matrix (Tüdös *et al.*, 1990) which is based on the 'neighborhood selectivity' (NS) of amino acids pairs (up to 10 residues distant from each other in the sequence) that characterizes whether a certain amino acid pair is favored or disfavored in terms of its observed frequency versus its expected frequency by chance. NS values were calculated over a large set of protein sequences ($\sim 2\times10^7$) derived from GenBank. The RReM matrix is a measure of the similarity of the NS values of the various amino acids to each other. The RReM matrix is found on a separate branch in a recent cluster analysis of published residue substitution tables, and is most closely related to various hydrophobicity measures (Tomii and Kanehisa, 1996).

In the DAS method hits are marked on the dot-plot surface at a very low cut-off, by default 1 standard deviation (SDU). For unrelated membrane proteins, the hits are unevenly distributed with the highest density of hits at the intersections of the transmembrane segments. This results in a chess-board like

Table I. Pairwise correlation coefficients of the various profiles for each protein sequence separately

SwissProt code	I	II	III	IV
ALKB_PSEOL.sw	0.7378	0.7079	0.6282	0.7310
ATPL_ECOLI.sw	0.8882	0.7164	0.8079	0.6950
COX2_PARDE.sw	0.6824	0.4256	0.7347	0.4276
COX3_PARDE.sw	0.5012	0.3881	0.5193	0.5711
CX1B_PARDE.sw	0.6557	0.6381	0.6115	0.6251
CYDA_ECOLI.sw	0.7448	0.6137	0.7746	0.6956
CYDB_ECOLI.sw	0.8190	0.6800	0.7893	0.7822
CYOA_ECOLI.sw	0.7195	0.5193	0.5557	0.5946
CYOB_ECOLI.sw	0.7207	0.5895	0.6543	0.6980
CYOC_ECOLI.sw	0.7052	0.6534	0.7536	0.6706
CYOD_ECOLI.sw	0.8388	0.7452	0.5583	0.7459
CYOE_ECOLI.sw	0.5645	0.4231	-0.0039	0.3687
DHG_ECOLI.sw	0.8208	0.5279	0.7768	0.7280
DMSC_ECOLI.sw	0.7283	0.6419	0.7308	0.7188
DSBB_ECOLI.sw	0.7001	0.6018	0.6747	0.6385
ENVZ_ECOLI.sw	0.8137	0.5625	0.8658	0.9603
EXBB_ECOLI.sw	0.8265	0.6134	0.8421	0.8920
EXBD_ECOLI.sw	0.8509	0.6578	0.8726	0.9145
FTSH_ECOLI.sw	0.7794	0.5424	n.a.	0.7503
FTSL_ECOLI.sw	0.8893	0.7834	0.9394	0.9711
FUCP_ECOLI.sw	0.6302	0.7101	0.6444	0.8972
GLPT_ECOLI.sw	0.6941	0.6888	0.6910	0.8576
HISM_SALTY.sw	0.6385	0.6366	0.5792	0.5696
HISQ_SALTY.sw	0.6995	0.3838	0.5179	0.6077
HOXN_ALCEU.sw	0.6042	0.5686	0.5152	0.5205
IMMA_CITFR.sw	0.7847	0.6530	0.7759	0.8311
KDPD_ECOLI.sw	0.7536	0.4873	0.8968	0.8051
KGTP_ECOLI.sw	0.6977	0.7086	0.7016	0.7657
LACY_ECOLI.sw	0.6902	0.6314	0.5858	0.7282
LSPA_ECOLI.sw	0.6799	0.5568	0.8061	0.7708
MALG_ECOLI.sw	0.6937	0.6950	0.7994	0.7098
MELB_ECOLI.sw	0.6473	0.6528	0.7532	0.7170
MOTA_ECOLI.sw	0.7143	0.6050	0.8529	0.9078
MOTB_ECOLI.sw	0.7109	0.5437	0.9245	0.8253
MTR_ECOLI.sw	0.6930	0.4595	0.6567	0.7162
OPPB_SALTY.sw	0.8031	0.6835	0.8185	0.9462
OPPC_SALTY.sw	0.7947	0.6574	0.7906	0.8040
PHOR_ECOLI.sw	0.7752	0.5252	0.7782	0.5434
RHAT_ECOLI.sw	0.6238	0.6738	0.6183	0.8655
SECD_ECOLI.sw	0.7765	0.6542	0.6953	0.6897
SECE_ECOLI.sw	0.7407	0.7092	0.6844	0.8951
SECY_ECOLI.sw	0.7468	0.6661	0.8448	0.8604
TOLQ_ECOLI.sw	0.8303	0.8204	0.7476	0.6993
TOLR_ECOLI.sw	0.8154	0.5171	0.8242	0.7450
average	0.7312	0.6134	0.7113	0.7373

I, experimental versus DAS; II, experimental versus <H>; III, experimental versus PHDhtm; IV, experimental versus TOPPRED

pattern. The DAS plot of two arbitrary chosen proteins of the database is shown in Figure 1. By summation, a 'cross-weighted cumulative score' profile (Cserzö *et al.*, 1994) can be calculated for each of the two proteins, with the transmembrane segments appearing as peaks.

A global DAS profile was calculated for each of the test proteins by averaging the 43 individual cumulative score profiles obtained for pairwise alignments with the other test set proteins. The RReM scoring matrix and default parameters (window 10 residues, cut-off 1.0 SDU) were used.

Since the DAS method compares each transmembrane segment against all the other transmembrane segments in the test set, the global DAS profile is very insensitive to the proteins included in the test set. In fact, the cross-weighted cumulative score profile for a given pair of sequences is in most cases almost identical to the global DAS profile, as illustrated in Figure 1.

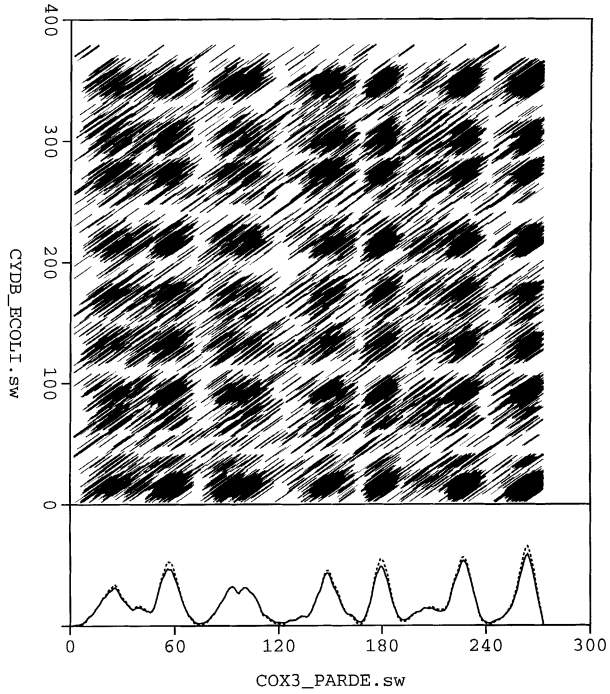

Fig. 1. DAS plot of two arbitrarily chosen proteins (COX3_PARDE versus CYDB_ECOLI). The cross weighted cumulative score profile (dotted line) and the global DAS profile (continuous line) calculated as the average of the cumulative score profiles obtained for comparisons with the other 43 proteins in the test set are also shown for COX3_PARDE. COX3_PARDE has seven and CYDB_ECOLI has eight transmembrane segments.

Reference predictions

For reference, standard hydrophobicity profiles were calculated for the proteins in the test set. A sliding window averaging with a trapezoid window was used (von Heijne, 1992). The window core size was 9 and the full size 11 to match the window size used in the DAS profile calculation. The Engelman–Steitz hydrophobicity scale was used (Engelman *et al.*, 1986). These profiles are referred to as <H> profiles.

Transmembrane helices were predicted using the TOPPRED algorithm (core window, 11 residues; full window, 21 residues), where the distribution of positively charged residues in the loops connecting the transmembrane helices is optimized in the final prediction (von Heijne, 1992). These predictions were transformed to topology profiles by setting a value of 1 for the predicted transmembrane regions and 0 for the rest. These profiles are referred to as TOPPRED profiles.

Finally, transmembrane helices were predicted using the PHDhtm server (Rost *et al.*, 1995) at http://www.embl-heidelberg.de/predictprotein, a program based on a trained neural network and multiply aligned test sequences. From the predicted topologies, PHDhtm profiles with a value of 1 for predicted transmembrane regions and 0 elsewhere were generated.

Statistical tests

To convert the DAS and <H> profiles into predictions of transmembrane segments, they were transformed so that the values were set to 1 where the original profiles were above a certain cut-off and to 0 in the rest of the sequence. This transformation resulted in square shaped profiles similar to the experimental, TOPPRED and PHDhtm profiles. The similarities of the different profiles (experimental, DAS, <H>,

TOPPRED and PHDhtm) were measured by their pairwise correlation coefficients calculated for each protein separately as well as for all the profiles together. In the latter test, the 44 profiles were concatenated and treated as a single, long profile.

The experimental determination of the ends of the transmembrane segments is uncertain and it is thus questionable to base the evaluation of the different methods on single-residue prediction performances. To minimize this problem the number of predicted transmembrane segments for a protein and the number of transmembrane segments overlapping with an experimental transmembrane segment were counted. The actual length of the overlapping portion was ignored. These tests are not sensitive to the uncertainties in the location of the transmembrane segments. The efficiency of the predictions were measured in terms of the following two ratios:

$$M = E_m/E_t \quad 1$$

$$C = P_m/P_t \quad 2$$

where E_m is the number of experimental transmembrane segments that overlap with a predicted transmembrane segment, E_t is the total number of experimental transmembrane segments, P_m is the number of predicted transmembrane segments that overlap with an experimental transmembrane segment and P_t is the total number of predicted transmembrane segments. E_m is not equal to P_m as in some cases the matching peaks are split and the same experimental transmembrane segment is thereby matched twice. There are also some examples of predictions when the same predicted peak matches two experimental transmembrane segments.

Results and discussion

To measure the similarities of the experimental versus predicted topology profiles, correlation coefficients were calculated between them for each sequence separately, Table I. The average correlations are 0.73, 0.61, 0.74 and 0.71 for the DAS, <H>, TOPPRED and PHDhtm profiles, respectively (for FTSH_ECOLI the neural network did not predict any transmembrane segment, thus the correlation coefficient is not applicable in this case. The corresponding values for DAS, <H> and TOPPRED were ignored in the averaging.)

DAS profiles are on average better correlated with the experimental profiles than are the <H> profiles. In only two cases is the <H> profile better correlated with the experimental profiles than the corresponding DAS profile.

The correlation between the PHDhtm and TOPPRED profiles and the experimental ones behaves differently. For most of the proteins these profiles are as good as the DAS profile, however, in a few cases they are much better while in a few other cases they are much worse. We could not find any correlation between the number of aligned sequences or their percentage identity to the query used in the multiple alignment and the accuracy of the PHDhtm prediction compared with the accuracy of the DAS method (data not shown).

In the next step the DAS and <H> profiles were transformed into square shaped prediction profiles as described in Materials and methods. The correlation of the transformed profiles with the concatenated experimental profile (see Materials and methods) as a function of the applied cut-off is shown in Figure 2. At low cut-off values transmembrane segments are predicted everywhere resulting in an elevated number of false predictions, whereas at high cut-offs the predictions miss the real peaks. This behavior results in optimal correlation

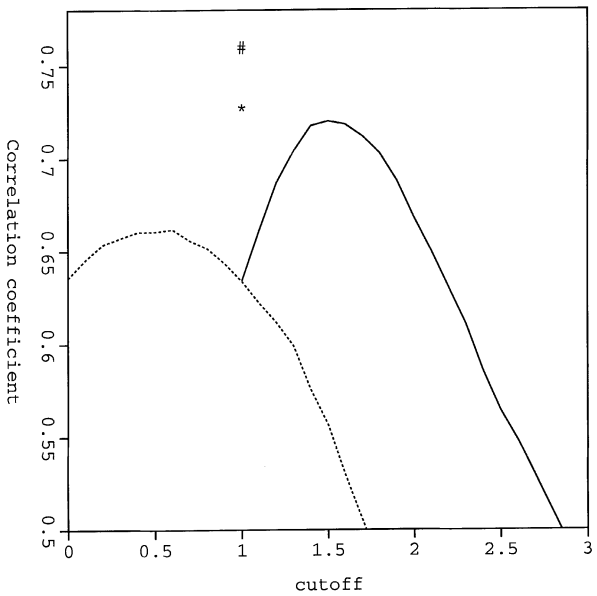

Fig. 2. Correlation coefficients of the transformed DAS (continuous line) and <H> (dotted line) profiles against the experimental ones as the function of the applied cut-off of the transformation. The corresponding value for the PHDhtm and TOPPRED profiles are marked by '*' and '#'.

coefficients of 0.66 and 0.72 at cut-off 0.6 and 1.5 for the <H> and the DAS profiles, respectively (the actual position of the optimum reflects only to the different scaling of the two types of profiles). The PHDhtm and TOPPRED profiles are optimized from the beginning, and their correlation coefficients in this test are 0.73 and 0.76, respectively.

The ends of the transmembrane segments listed in SwissProt are uncertain, and the limited precision of the database is thus mixed up with the limitation of the applied methods. To address this problem, only the number of predicted segments and the number of segments overlapping with an experimental segment were counted. The actual length of the overlapping portion was ignored. The efficiency of the prediction at a given cut-off was measured by the ratios *M* and *C* (Materials and methods, Eqns 1 and 2). *M* decreases and *C* increases with the cut-off. The behavior of *M* and *C* as a function of the cut-off is shown in Figure 3 for the DAS and <H> profiles. The optimal cut-off is found around the intersection of the *M* and *C* curves. The efficiency of the predictions in terms of the probability to match a real peak (*M*) and the probability that a predicted peak is a real one (*C*) are given in Table II. The geometric mean of these values are also presented in the last column to characterize the overall predictive power of the methods. Again, the DAS method is found to perform better than <H>, on par with PHDhtm and slightly worse than TOPPRED.

A common problem in all methods that try to predict transmembrane segments is that closely spaced pairs of segments sometimes show up as a single, wide peak in the prediction output, and, conversely, that a single transmembrane segment sometimes shows up as a pair of closely spaced, narrow peaks. To address this problem, one usually resorts to ad hoc rules that are found to increase the method's performance (Rost *et al.*, 1995, 1996). However, given that the number of erroneous predictions on the present test set is small from the outset, the statistical significance of these extra rules is often doubtful and one always runs the risk of over-fitting the rules to a few, special cases that happen to be present.

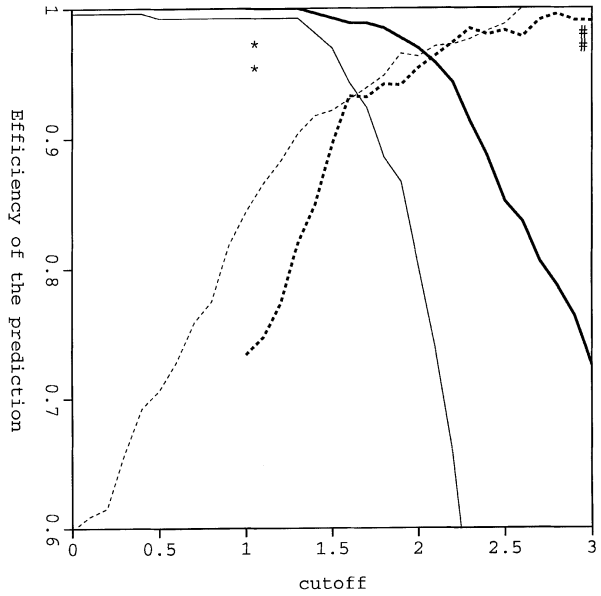

Fig. 3. Ratio of correct matches (*M*, continuous lines) and ratio of correct predictions (*C*, dotted lines) as the function of the applied cut-off of the transformation for the DAS (thick lines) and <H> (thin lines) profiles. The corresponding values for the PHDhtm and TOPPRED profiles are marked by '*' and '#'.

Table II. Optimal *M* and *C* ratios

Profile (cut-off)	*M*	*C*	$\sqrt{M * C}$
<H> (1.5)	0.969	0.921	0.945
DAS (2.2)	0.943	0.974	0.958
PHDhtm	0.950	0.969	0.959
TOPPRED	0.981	0.971	0.976

The geometric mean $\sqrt{M * C}$ is listed in the last column as the measure of the overall predictive power.

We have found that the DAS method can be further improved in this way (the best values for *C* and *M* obtained so far are $C = 0.969$ and $M = 0.958$; data not shown), though we do not consider this improvement sufficiently large to feel confident that the extra complexity introduced is really warranted. In practice, it is probably better that the person using the method assesses the results with these possible complications in mind than to pre-process the output using rules of questionable value.

An interesting aspect of the analysis is the problem of the precision and fidelity of the database. In the SwissProt feature tables for the test set sequences there are nine transmembrane segments not mentioned while two are incorrectly located, corresponding to an error rate of 4.2%. As the prediction methods typically are optimized to match the transmembrane segments given in SwissProt, they are sensitive to errors in the database itself. On the other hand the 'RReM' matrix—which is responsible for the sensitivity of transmembrane detection of the DAS method—was not optimized for transmembrane protein predictions but was derived from 'neighborhood selectivity' data over a large set of proteins that reflects how the different amino acids prefer the others in their sequential neighborhood. It is thus not directly related to the hydrophobic properties of amino acids in transmembrane segments, although it correlates reasonably well with hydrophobicity indices (Tomii and Kanehisa, 1996). As a consequence we might consider the DAS profiles as relatively independent from errors in the database.

In conclusion, we have compared the performance of four different transmembrane segment prediction methods: a sliding window averaging with trapezoid window (<H>), a method (TOPPRED) based on the 'positive inside' rule (von Heijne, 1992), a neural network method (PHDhtm) including information from multiply aligned sequences (Rost *et al*., 1995, 1996) and the new DAS method. The predictive power of DAS and PHDhtm is essentially the same while the single-sequence based <H> method performs significantly worse when applied to a test set of 44 well characterized prokaryotic membrane proteins. Incorporating extra information related to the positive inside rule (TOPPRED) brings the predictive power to the level of the two other methods. This suggests that the DAS method, which uses only single sequence information, is on par with the PHDhtm method (which uses multiple sequence alignments) and TOPPRED (which uses extra information in the form of the distribution of positively charged residues) in predicting transmembrane segments in prokaryotic inner membrane proteins.

A WWW server running the DAS algorithm is available at http://www.biokemi.su.se/~server/DAS/.

Acknowledgements

This work was supported through the cooperation framework of the Royal Swedish Academy of Sciences and the Hungarian Academy of Sciences, and by grants from the Swedish Technical Sciences Research Council (TFR) and the Magnus Bergvall foundation to A.E., and from the Swedish Natural Sciences Research Council to G.v.H.

References

Bairoch,A. and Boeckmann,B. (1991) *Nucleic Acids Res.*, **19**, 2247–2249.
Cserzö,M., Bernassau,J.M., Simon,I. and Maigret,B. (1994) *J. Mol. Biol.*, **243**, 388–396.
Engelman,D.M., Steitz,T.A. and Goldman,A. (1986) *Annu. Rev. Biophys. Chem.*, **15**, 321–353.
Gött,P. and Boos,W. (1988) *Mol. Microbiol.*, **2**, 655–663.
Jones,D.T., Taylor,W.R. and Thornton,J.M. (1994) *Biochemistry*, **33**, 3038–3049.
Kampfenkel,K. and Braun,V. (1993) *J. Bacteriol.*, **175**, 4485–4491.
Persson,B. and Argos,P. (1996) *Protein Sci.*, **5**, 363–371.
Pogliano,K.J. and Beckwith,J. (1994) *J. Bacteriol.*, **176**, 804–814.
Rost,B., Casadio,R., Fariselli,P. and Sander,C. (1995) *Protein Sci.*, **4**, 521–533.
Rost,B., Fariselli,P. and Casadio,R. (1996) *Protein Sci.*, **5**, 1704–1718.
Tomii,K. and Kanehisa,M. (1996) *Protein Engng*, **9**, 27–36.
Tüdös,E., Cserzö,M. and Simon,I. (1990) *Int. J. Peptide Protein Res.*, **36**, 236–239.
Vianney,A., Lewin,T.M., Beyer,W.F., Lazzaroni,J.C., Portalier,R. and Webster,R.E. (1994) *J. Bacteriol.*, **176**, 822–829.
von Heijne,G. (1992) *J. Mol. Biol.*, **225**, 487–494.
von Heijne,G. (1994) *Annu. Rev. Biophys. Biomol. Struct.*, **23**, 167–192.
Zimmann,P., Puppe,W. and Altendorf,K. (1995) *J. Biol. Chem.*, **270**, 28282–28288.

Received 30 October 1996; revised February 17, 1997; accepted March 4, 1997