The Use of Receiver Operating Characteristic (ROC) Analysis to Evaluate Sequence Matching

Michael Gribskov[1] & Nina L. Robinson[2]

[1] Corresponding author:
San Diego Supercomputer Center
P.O. Box 85608
San Diego CA 92186-9784
(619) 534 - 8312

UPS/FedEx/Courier Address:
10100 Hopkins Dr.
La Jolla 92093-0505
FAX: (619) 534 - 5152

[2] Sequana Therapeutics, Inc.
11099 North Torrey Pines Road
Suite 160
La Jolla CA 92037-1029
(619) 452 - 6550

FAX: (619) 452 - 6653

Abstract

In this paper, we borrow the idea of the Receiver Operating Characteristic (ROC) from clinical medicine and demonstrate its application to sequence comparison. The ROC includes elements of both sensitivity and specificity, and is a quantitative measure of the usefulness of a diagnostic. The ROC is used in this work to investigate the effects of scoring table and gap penalties on database searches. Studies on three families of proteins, 4Fe-4S ferredoxins, lysR bacterial regulatory proteins, and bacterial RNA polymerase sigma factors lead to the following conclusions: Sequence families are quite idiosyncratic, but the best PAM distance for database searches using the Smith-Waterman method is somewhat larger than predicted by theoretical methods, about 200 PAM. The length independent gap penalty (gap initiation penalty) is quite important, but shows a broad peak at values of about 20 - 24. The length dependent gap penalty (gap extension penalty) is almost irrelevant suggesting that successful database searches rely only to a limited degree on gapped alignments. Taken together, these observations lead to the conclusion that the optimal conditions for alignments and database searches are not, and should not be expected to be, the same.

Introduction

The sensitivity and specificity of sequence matching techniques are often evaluated by subjective methods (Pearson, 1991; Henikoff & Henikoff, 1992). Historically, this can be attributed to both the difficulty of performing large numbers of comparisons and to a lack of quantitative methods for making comparisons. With the growing use of sequence comparisons as a basis for the automatic annotation of macromolecular sequence databases, it is increasingly important to evaluate the statistical performance of these methods under various conditions.

In this paper we examine the use of the receiver operating characteristic (ROC) to evaluate the sensitivity and specificity of sequence comparisons. ROC analysis has long been used in clinical applications to evaluate the usefulness of diagnostic tests and to determine proper cutoff thresholds (for a recent review, see Zweig & Campbell, 1993). A common clinical application is the evaluation of the usefulness of a test, for instance serum cholesterol level, as a diagnostic for a medical condition such as heart disease. This is very similar to the question being evaluated in automatic annotation of sequence databases: "given a similarity score for comparison to a particular sequence (diagnostic test), what is the probability that the sequences are homologous." ROC analysis offers a systematic and well understood way to evaluate the effect of algorithmic parameters, scoring systems, and other factors on sequence comparisons.

Methods

ROC Analysis

The ROC is evaluated by means of a plot of the true positive fraction (sensitivity) versus the true negative fraction (1-specificity) using a continuously varying decision threshold. In practice, the plot is produced by classifying each datapoint as positive or negative according to outcome (i.e., whether the sequences are homologous or not homologous), and plotting the points in rank order of the diagnostic measure (i.e., alignment similarity score). More simply, each observation describes a point on a two dimensional plot where the ordinate indicates the fraction of positives

with an equal or greater score, and the abscissa indicates fraction of negatives with an equal or greater score (fig 1). A diagnostic with good discrimination will rank the true positives above the true negatives and therefore will have a true positive fraction greater or equal to the true negative fraction at all points. The ROC curve in this case will lie in the upper left corner (fig1a). A diagnostic with no discrimination will have the positives and negatives mixed together and will produce a line with a slope of 1 (fig 1b).

The area under the ROC curve measures the probability of correct classification (formally, the area under the ROC curve has been shown to be equivalent to the value of a Mann-Whitney or Wilcoxon non-parametric test, Bamber, 1975). An area of 0.9, for instance, indicates that a sequence chosen from the positive (truly homologous) group has a probability of 0.9 of scoring higher than a sequence chosen from the negative group. The area under the ROC curve can be used without modification to examine the sensitivity and specificity of sequence comparison methods. However a slight modification is useful in the context of macromolecular sequences and database searching.

Even the largest sequence families have only a few hundred members. This means that in a typical database search, the true negatives will be present in hundreds to thousands of fold excess. The larger number of true negatives is not an intrinsic difficulty for the method, however since most sequences are fairly good discriminators for their respective families, the area under ROC curves plotted over the entire database will usually be very close to one. The relative sensitivity and specificity of various methods can still be evaluated by comparison of the areas under the respective curves but the areas must be calculated to 5 or 6 significant figures. Furthermore, calculation of the ROC curve over the full extent of the true negative results requires that all scores be written out and saved - a large volume of results when thousands of conditions are to be compared. More importantly, the results of database searches are seldom examined beyond the point where several tens of true negative results are observed. Therefore from a practical viewpoint, true positive results that have scores lower than a page or so of true negative results are unobserved and might as well be considered to have a score of zero. For these reasons we define

ROC_{50} as follows: The ROC_{50} is the area under the ROC curve plotted until 50 true negatives are found. The number of negatives is thus fixed at 50. ROC_{50} is simpler to calculate than the complete area under an ROC curve and can be expressed in fewer significant digits. It also takes into account the observation that true positives that occur after roughly a page worth (about 50) of true negatives are essentially not observed.

Evaluation of Sequence Similarity

Sequence comparisons were performed using a local similarity dynamic programming algorithm (Smith & Waterman, 1981), as implemented in the MPSRCH program (IntelliGenetics, Inc., Mountain View, CA). Results were normalized for systematic dependence on sequence length using the "ranking function" method implemented in this package. MPSRCH was used in the affine gap scoring mode in which a length independent (gap initiation) and a length dependent (gap extension) penalty are used. Each family of sequences was evaluated using four PAM scoring tables in the range 50 - 250 PAM, and with a two dimensional grid of gap opening and gap extension parameter sampled at an interval of two. Generally about 12 settings of both the gap initiation parameter and the gap extension parameter were investigated for each query sequence. Approximately 7000 database searches were performed in the work reported here.

BLAST database searches (Altschul et al., 1990) were performed by e-mail using BLAST version 1.4. Each query sequence was compared to all five scoring systems available to the e-mail service: IDENTITY, PAM40, PAM100, PAM250, and BLOSUM62. Searches were run using default parameters except that the EXPECT value was set to 200 in order to retrieve enough scores for ROC_{50} analysis.

Identification of True Positives

True positives were identified by a combination of criteria. The initial group of sequences for each family was determined based on sequence annotation and the PROSITE database (Bairoch & Boeckmann, 1993). Five to 15 sequences from this original group were then used as queries in database searches using BLAST and MPSRCH and the results of the searches examined for

further distantly related sequences, i.e., sequences with significant or high scores with all of the query sequences. Sequences with comparison scores that were high but below the normal threshold of significance were further investigated. These candidate sequences were accepted as part of the group when database searches using the candidate sequences as a query were found to detect the group with an ROC_{50} greater than 0.2. The results of this cyclic refinement of the true positive group are in good agreement with published work on the sequence families studied here (Bairoch & Boeckmann, 1993;Schell, 1993; Lonetto et al, 1994). Small errors in the identification of the true positive group will effect the ROC_{50} statistic, but only in an absolute sense. If an actual positive sequence is omitted from the defined positive group, all queries will be equivalently affected by the presence of a high scoring (apparent) negative in the ROC plot. The presence of a small number of misidentified sequences should not greatly affect the optimal searching conditions since these depend on the separation of the entire positive and negative groups.

Results

Three protein families have been selected as examples of the usefulness of ROC_{50} analysis. The families discussed below were selected because of the differing pattern of conservation within each family. The ferredoxins contain a short highly conserved "signature-like" sequence rich in rare amino acid residues. The lysR family consists of an entire domain that is combined with various kinds of domains in different members of the family. Finally, the bacterial sigma factors represent a diverse group in which the conserved regions are separated by unconserved regions with very variable lengths and compositions. These families have been chosen to cover a wide range of types of families, but obviously do not represent the comprehensive sample needed to firmly establish database searching conditions.

Ferredoxins

The 4Fe-4S ferredoxins are small proteins involved in electron transfer. Ferredoxin like molecules also function in photosynthesis, and are found as independent domains in a variety of

enzymes involved in oxidation-reduction reactions. These ferredoxins bind a 4Fe-4S cluster at a highly conserved 26 residue sequence containing the four cysteines that bind the Fe. Most of the members of this family have two copies of the characteristic 26 residue repeat and therefore bear two 4Fe-4S centers. Figure 2a shows that the optimal PAM distance varies with the probe sequence, but is generally about 150 to 200 PAM. Figure 2b shows the best ROC_{50} obtained at different values of the gap initiation penalty. This is clearly an important variable, but shows a broad peak in the 16 - 26 range. Finally figure 2c shows the best ROC_{50} obtained for various settings of the gap extension parameter. This plot is striking in its flatness, implying that this parameter is essentially irrelevant.

LysR family

The lysR family is a group of bacterial transcriptional regulatory proteins (Henikoff et al., 1988; Schell, 1993). These proteins share a conserved amino-terminal domain that appears to bind to DNA via a "helix-turn-helix" motif. The sequence similarity in this family comprises the entire DNA binding domain which spans roughly 65 - 100 residues. The carboxyl-terminal domain of these proteins is heterologous, i.e., the DNA binding domain in found fused to a number of distinct types of C-terminal domains. Correct alignment of the DNA binding domain should require a number of gaps. Figure 2d-f show the results of the parameter screening for this group of sequences. Again there is some variation in the optimal PAM scoring table depending on the probe sequence, and a clear dependence on the gap initiation parameter. The plot (fig 2f) for the gap extension penalty is again nearly flat suggesting the relative irrelevance of this parameter.

Sigma factors

Bacterial sigma factors are a widespread group of sequences with a distinctive pattern of conserved and variable regions. The proteins vary greatly in length, from approximately 280 residues to more than 600. This variation in length is due to both variation in the length of the variable regions and to the lack of some of the conserved regions in the smaller proteins. Figure 2g-i show the results of the parameter screening for this group of sequences. Again there is some

variation in the optimal PAM scoring table to use depending on the probe sequence, and a clear dependence on the gap initiation parameter. Figure 2i shows a more pronounced dependence of the best ROC_{50} on the gap extension penalty than in the previous two groups. However, the effect is still relatively minor, amounting to a difference of approximately 0.03 in the ROC_{50}. This amount of difference corresponds to roughly 1 sequence difference in sensitivity.

Comparison to BLAST

The results for the ferredoxins, lysrR family and sigma factors suggest that gap penalties are relatively irrelevant as long as they are high enough to suppress matches to unrelated sequences. The lack of dependence on the gap extension penalty, in particular, suggests that a method that does not consider gaps might be just as efficient in database searches as the Smith-Waterman method. To investigate this possibility, the query sequences used above were also used as queries for BLAST searches (Altschul et al., 1990) with the results shown in Table 1.

TABLE 1: Comparison of BLAST and MPSRCH[a]

	BLAST		MPSRCH			
Probe	**ROC_{50}**	**Matrix**	**ROC_{50}**	**Matrix**	**Open**	**Ext**
FER_ENTHI	0.812	PAM120	**0.951**	PAM150	24	2
FER_THETH	0.805	PAM120	**0.889**	PAM150	22	10
ASRA_SALTY	**0.521**	BLOSUM62	0.455	PAM100	22	22
DCMA_METSO	0.426	PAM250	**0.503**	PAM200	22	2
FRXB_WHEAT	**0.823**	BLOSUM62	0.811	PAM200	24	2
PSAC_CHLRE	0.803	BLOSUM62	**0.932**	PAM200	26	2
AMPR_RHOCA	0.935	PAM250	**0.972**	PAM200	16	4
BLAA_STRCI	0.690	BLOSUM62	**0.743**	PAM150	16	8
GLTC_BACSU	**0.891**	BLOSUM62	**0.891**	PAM200	18	18
METR_SALTY	0.951	BLOSUM62	**0.999**	PAM200	26	26
FLIA_SALTY	0.861	BLOSUM62	**0.939**	PAM100	20	2
RP32_ECOLI	0.851	PAM250	**0.959**	PAM150	14	2
RP70_ECOLI	0.852	PAM250	**0.956**	PAM250	26	2
RPSH_BACSU	0.740	BLOSUM62	**0.857**	PAM250	24	24
RPSK_BACSU	0.873	BLOSUM62	**0.944**	PAM100	14	2

[a] Shaded ROC_{50} values indicate equal or better discrimination.

While these results are too incomplete to show that the Smith-Waterman algorithm is superior to the one implemented in BLAST, the generally better ROC_{50} values seen with MPSRCH suggest

there is some benefit gained from taking gaps into consideration. It is particularly interesting to note that BLAST performs better than Smith-Waterman in two of six cases for the ferredoxins where the alignments should be short and without gaps. For the lysR family, where alignments should benefit from limited introduction of gaps, BLAST is equal to the Smith-Waterman approach in only one case. And in the case of sigma factor family, where the introduction of gaps should be most important, Smith-Waterman is better in all cases tested.

Discussion

The ROC_{50} statistic provides an efficient and quantitative way to evaluate sequence comparison methods. The ROC_{50} has the notable feature of insensitivity to threshold. Consider the example shown in figure 1. If one was to compare the distributions in figure 1a and figure 1b using a threshold diagnostic value of 1.0, no difference could be discerned. Similarly a threshold of 95% of the maximum diagnostic value (i.e., a value of 19.0) would not reveal the dramatic difference between distributions 1a and 1b. Furthermore, the ROC_{50} is sensitive to the order of the positive and negative observations, or equivalently to the specificity of the assay. Queries that have the same fraction of positives and negatives above a specified threshold may still be distinctly different in their ability to separate positives and negatives. This is detected in the ROC_{50} analysis but not with simple threshold analysis. The combination of both sensitivity and specificity into a single measure make the ROC_{50} especially convenient for analysing the results of database searches.

The results reported here are unexpected in several ways. Firstly, the optimal PAM distance for these sequence groups appears to be roughly 200 PAM rather than the 100-120 suggested by theoretical considerations (Altschul, 1991). Secondly, while the gap initiation penalty is very important in achieving the best results, the ROC_{50} plot for this parameter shows a rather broad peak. Reasonable values for this parameter are approximately 20 to 24. Thirdly, and surprisingly, the gap extension penalty is of relatively little importance. This lack of effect of a length

dependent parameter implies that little or no sensitivity is gained by allowing long gaps in the alignments between the query and true positive sequences.

The reason for this surprising behavior can be explained by the fundamental difference between the searching and alignment operations. When a sequence alignment is performed, two sequences that are presumed to be related are compared and the scoring system and penalties are adjusted to give the best alignment. In this context, a large component of the criterion for what is best is the length of the aligned region (assuming that the sequences in question are indeed homologous). In a database search however, related sequences are outnumbered by unrelated sequences by a factor of 1:100 to 1:10000. Under these conditions the sensitivity of the search becomes dominated by what happens with the unrelated sequences. Conditions that optimize the comparison of related sequences in the alignment sense, also greatly increase the scores for matching to an unrelated sequence. This behavior is very reasonable when one thinks about it, but leads to a conclusion that has not been widely appreciated: The optimal conditions for alignments and database searches are not, and should not be expected to be, the same.

This conclusion agrees with the speculation of Vingron and Waterman (1994) that good contrast in database searches might be obtained when the overall score depends on a few high scoring regions without gaps rather than the overall comparison. Such alignments are found in the "logarithmic region" of the alignment phase space where the gap penalties are high. This conjecture is supported by this work. Nevertheless, the results shown in Table 1 suggest that some allowance for gaps improves the ability of most queries to significantly match to homologous sequences and therefore provides a continuing rationale for the use of the Smith-Waterman algorithm in database searching.

Acknowledgments

This work was supported by the National Science Foundation through cooperative agreement ASC-8902825 with the San Diego Supercomputer Center and by the National Institutes of Health through award P41 RR08605. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views or policies of the National Science Foundation, the National Institutes of Health, other supporters of the San Diego Supercomputer Center, or Sequana Therapeutics, Inc. We thank IntelliGenetics for access to the MPSRCH software prior to its official release.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) *J.Mol.Biol.* **215,** 403-410.

Altschul, S.F. (1991) *J.Mol.Biol* **219**, 555-565.

Bairoch, A. & Boeckmann, B. (1993) *Nucleic Acids Res.* **21**, 3093-3096.

Bamber, D. (1975) *J.Math.Psych.* **12**, 387-415.

Henikoff, S., Haughn, G.W., Calvo, J.M. & Wallace, J.C. (1988) *Proc.Natl.Acad.Sci* **85**, 6602-6606.

Henikoff, S. & Henikoff, J. (1992) *Proc.Natl.Acad.Sci.* **89**, 10915-10919.

Lonetto, M, Gribskov, M., & Gross, C.A. (1992) *J.Bact.* **174**, 3843-3849.

Pearson, W.R. (1991) *Genomics* **11**, 635-50.

Schell, M.A. (1993) *Ann.Rev.Microbiol.* **47**, 597-626.

Smith, T.F. & Waterman M.S. (1981) *J.Mol.Biol.* **147**, 195-197.

Vingron, M. & Waterman, M.S. (1994) *J.Mol.Biol.* **235**, 1-12.

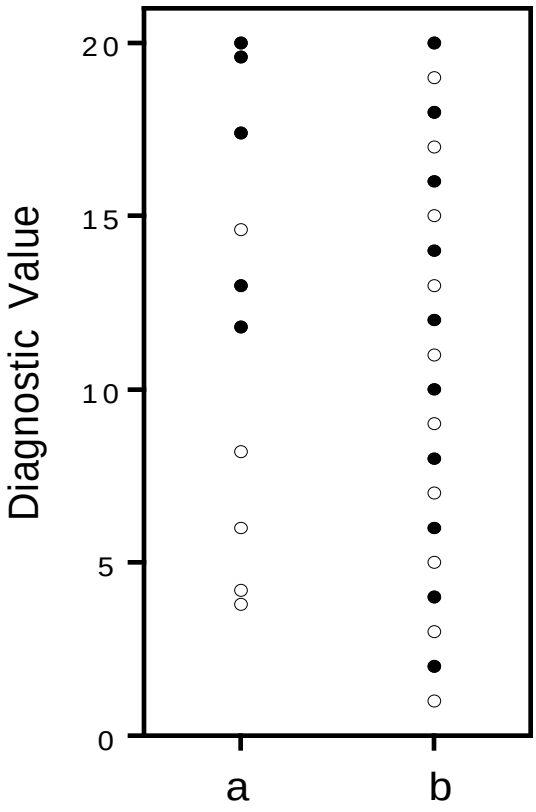
Zweig, M.H. & Campbell, G. (1993) *Clin.Chem.* **39**, 561-577.

Figure Legends

Figure 1. Construction of the ROC curve. The left panel shows two sets of datapoints. Set a represents data for a diagnostic with good discrimination, set b represents data for a diagnostic with no discrimination. Positive points are shown as filled circles (●), negative points as open circles (o). The right panel shows the ROC curves for the same sets of data with set a shown as filled circles (●) and set b shown as open circles (o).

Figure 2. Effect of scoring table and gap penalties on database searches. The left hand panel (a,d,g) shows the maximum ROC_{50} for any combination of gap initiation and gap extension penalties at each PAM distance. The center panel (b,e,h) shows the maximum ROC_{50} for any gap extension penalty at the optimal PAM distance. The right hand panel (c,f,i) shows the maximum ROC_{50} for any gap opening penalty at the optimal PAM distance. Fig 2a-2c show ROC_{50} curves examining the effect of PAM table, gap initiation penalty, and gap extension penalty on the ability of the following sequences to detect the 4Fe-4S family of ferredoxins: FER_ENTHI (open circle), FER_THETH (open square), ASRA_SALTY (diamond), DCMA_METSO (triangle), FRXB_WHEAT (filled circle), PSAC_CHLRE (filled square). Fig 2d-2f show similar curves examining the ability of the following sequences to detect the lysR family of transcriptional regulators: AMPR_RHOCA (circle), BLAA_STRCI (square), GLTC_BACSU (diamond), METR_SALTY(triangle). Fig 2g-2i show similar curves examining the ability of the following sequences to detect the bacterial RNA polymerase sigma factor family: FLIA_SALTY (open circle), RP32_ECOLI (square), RP70_ECOLI (diamond), RPSH_BACSU (triangle), RPSK_BACSU (filled circle).

Figure 1


Diagnostic Value
20
15
10
5
0
a
b

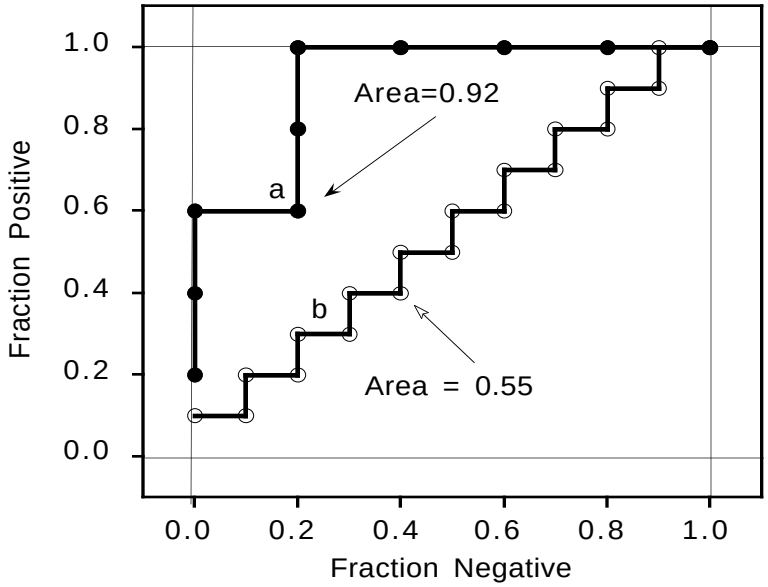

Fraction Positive
1.0
0.8
0.6
0.4
0.2
0.0
Area=0.92
a
b
Area = 0.55
0.0
0.2
0.4
0.6
0.8
1.0
Fraction Negative

Fig 2a

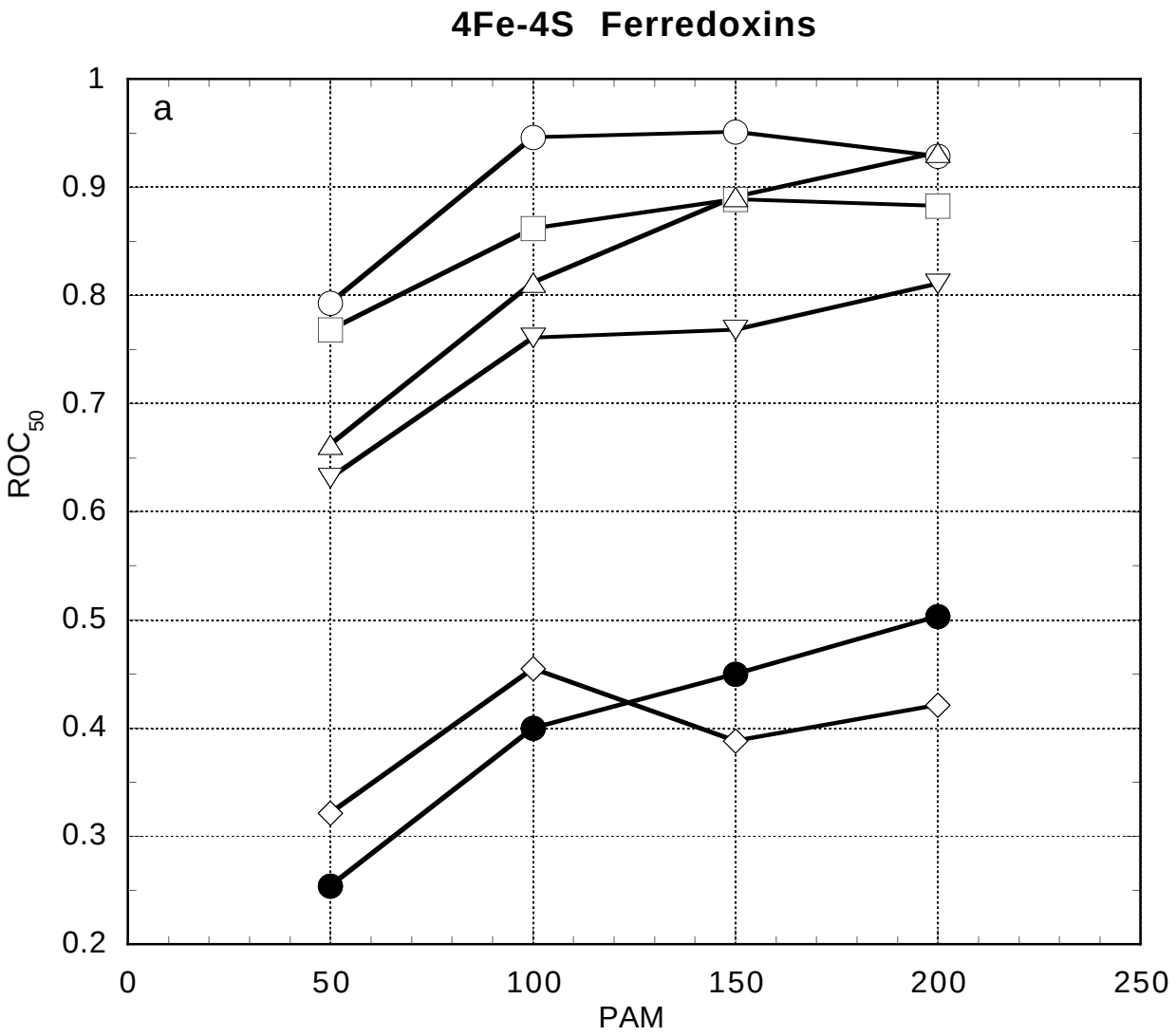

4Fe-4S Ferredoxins
a
ROC50
PAM
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
50
100
150
200
250

Fig 2b

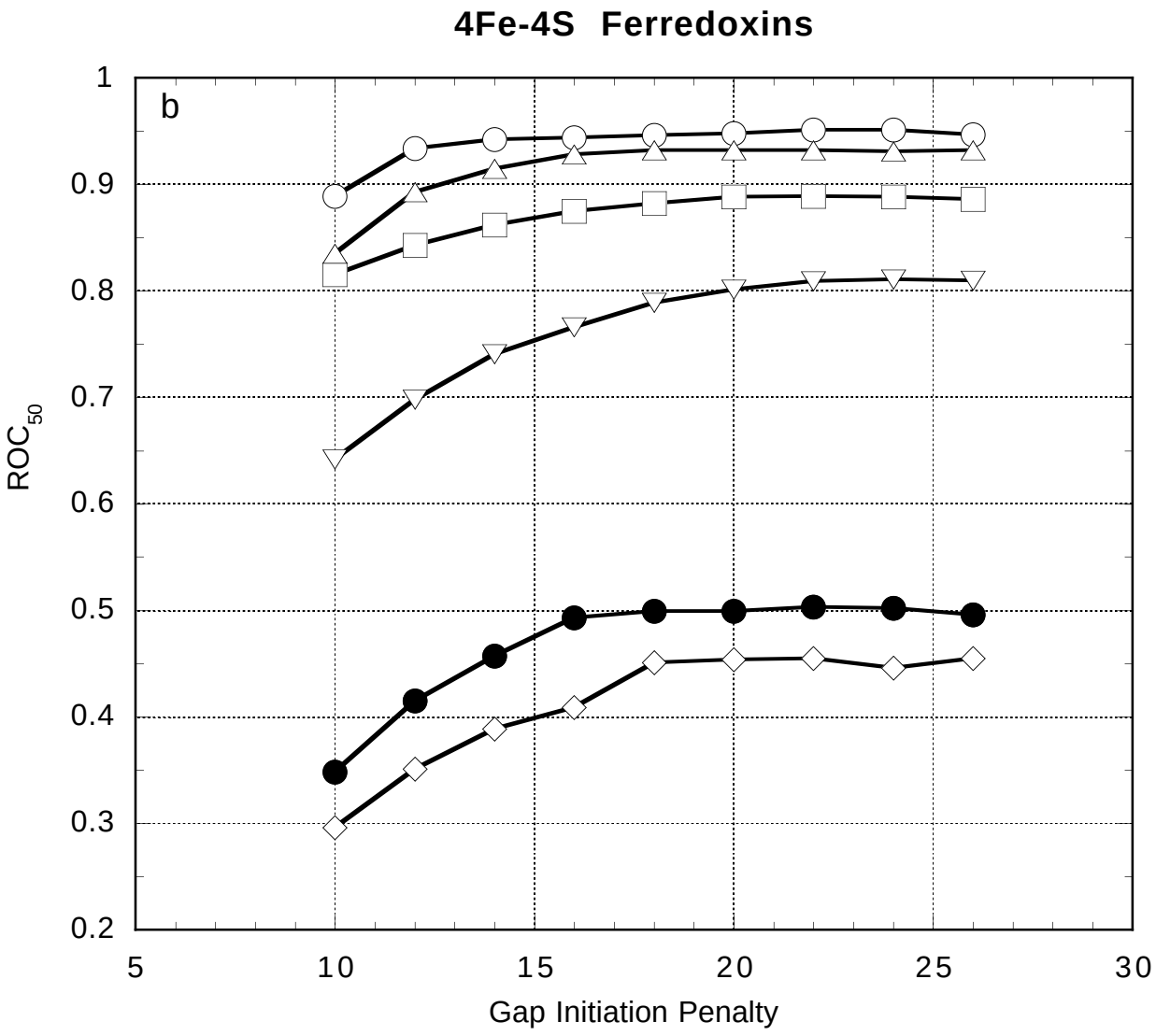

4Fe-4S Ferredoxins
b
ROC50
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
5
10
15
20
25
30
Gap Initiation Penalty

Fig 2c

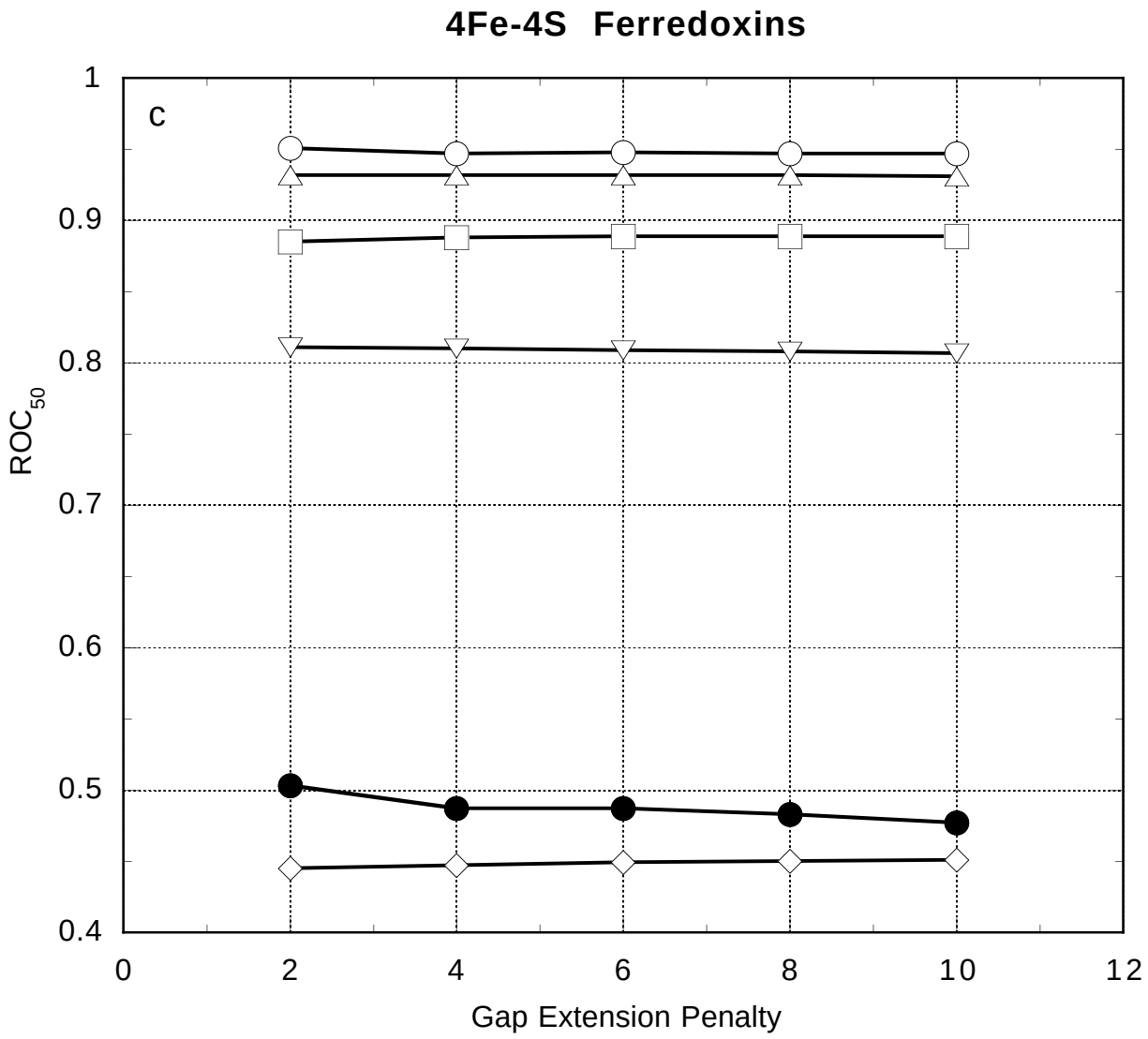

4Fe-4S Ferredoxins
c
1
0.9
0.8
0.7
0.6
0.5
0.4
ROC50
0
2
4
6
8
10
12
Gap Extension Penalty

Fig 2d

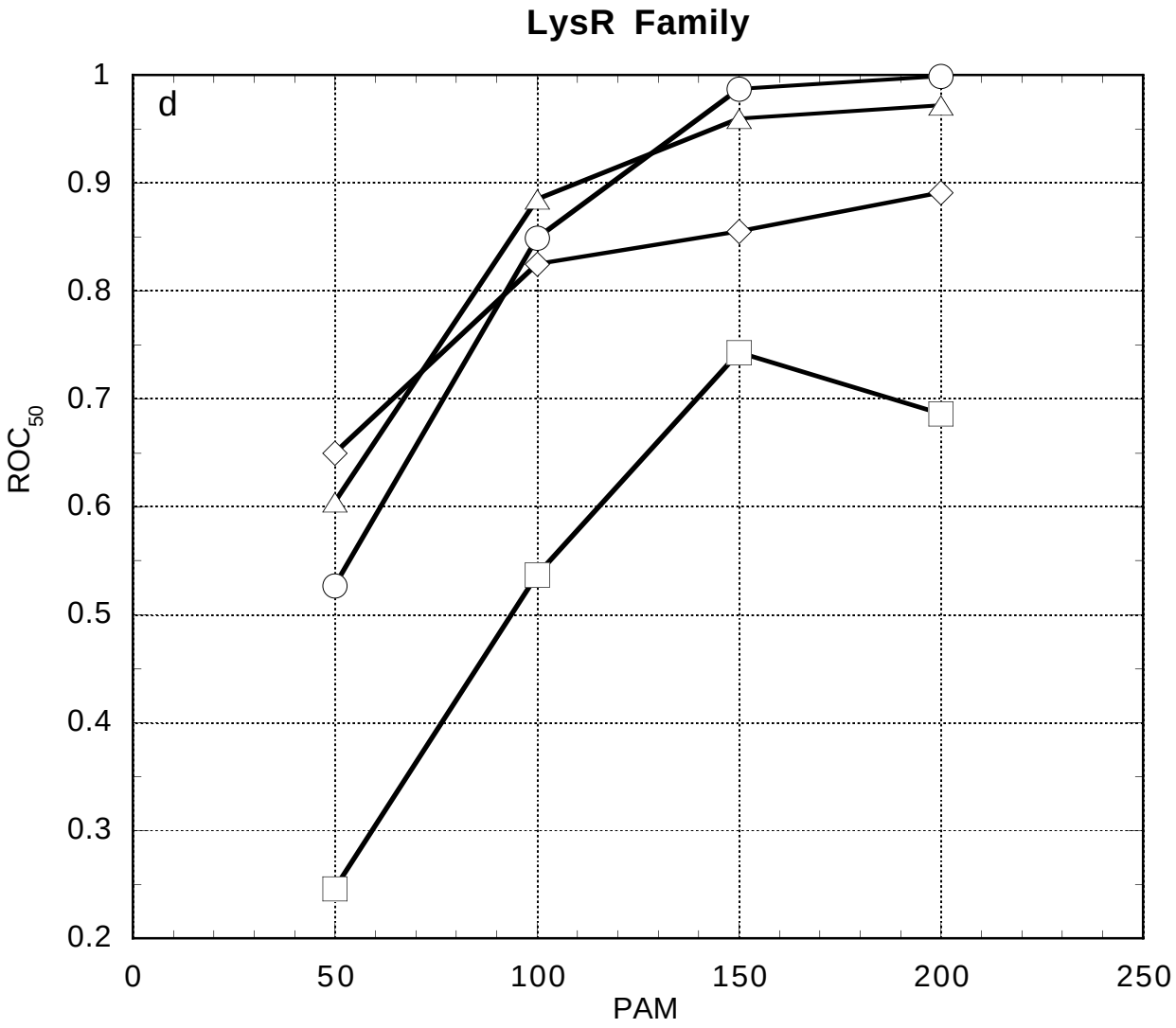

LysR Family
d
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
ROC50
0
50
100
150
200
250
PAM

Fig 2e

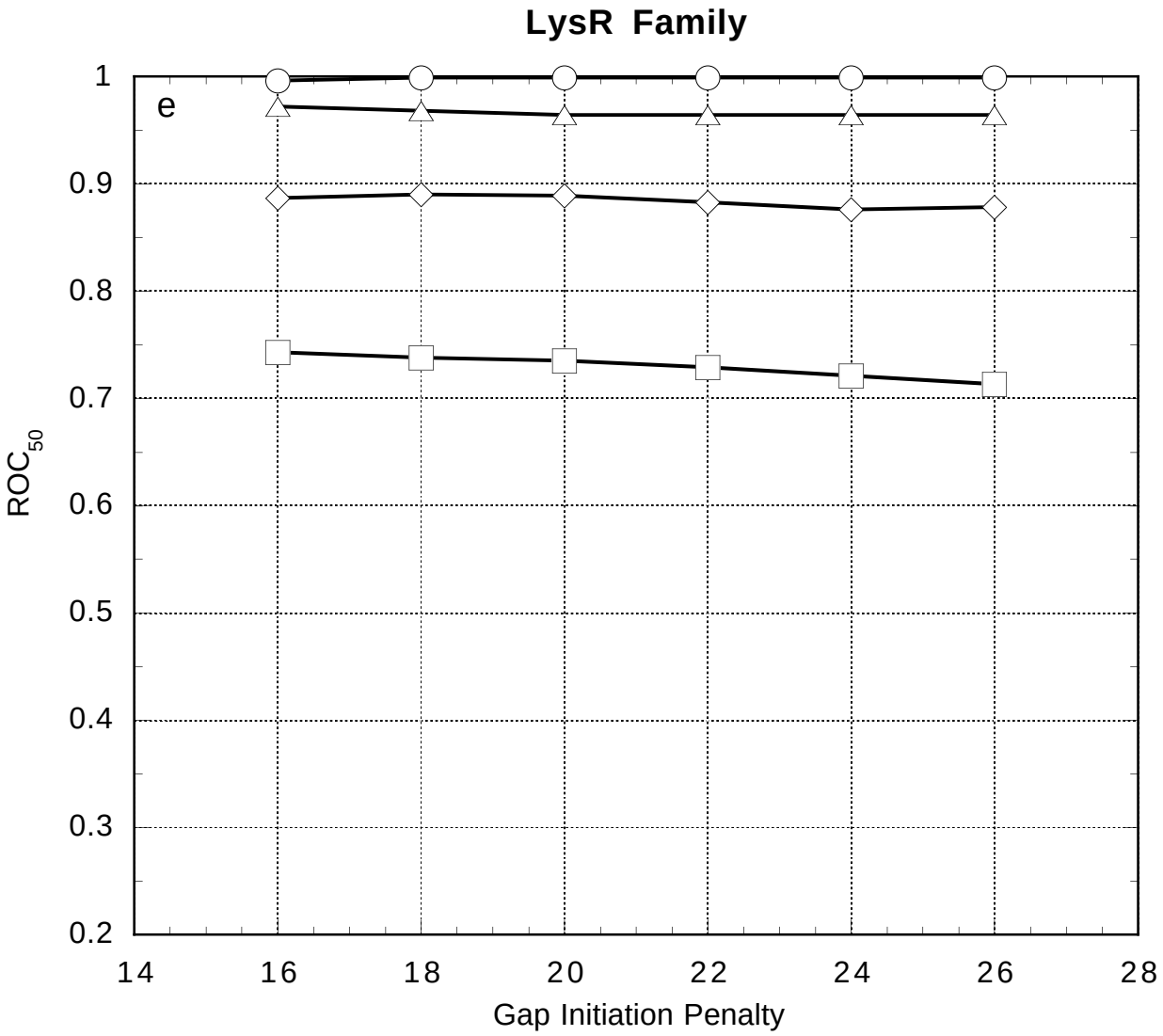

LysR Family
e
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
ROC50
14
16
18
20
22
24
26
28
Gap Initiation Penalty

Fig 2f

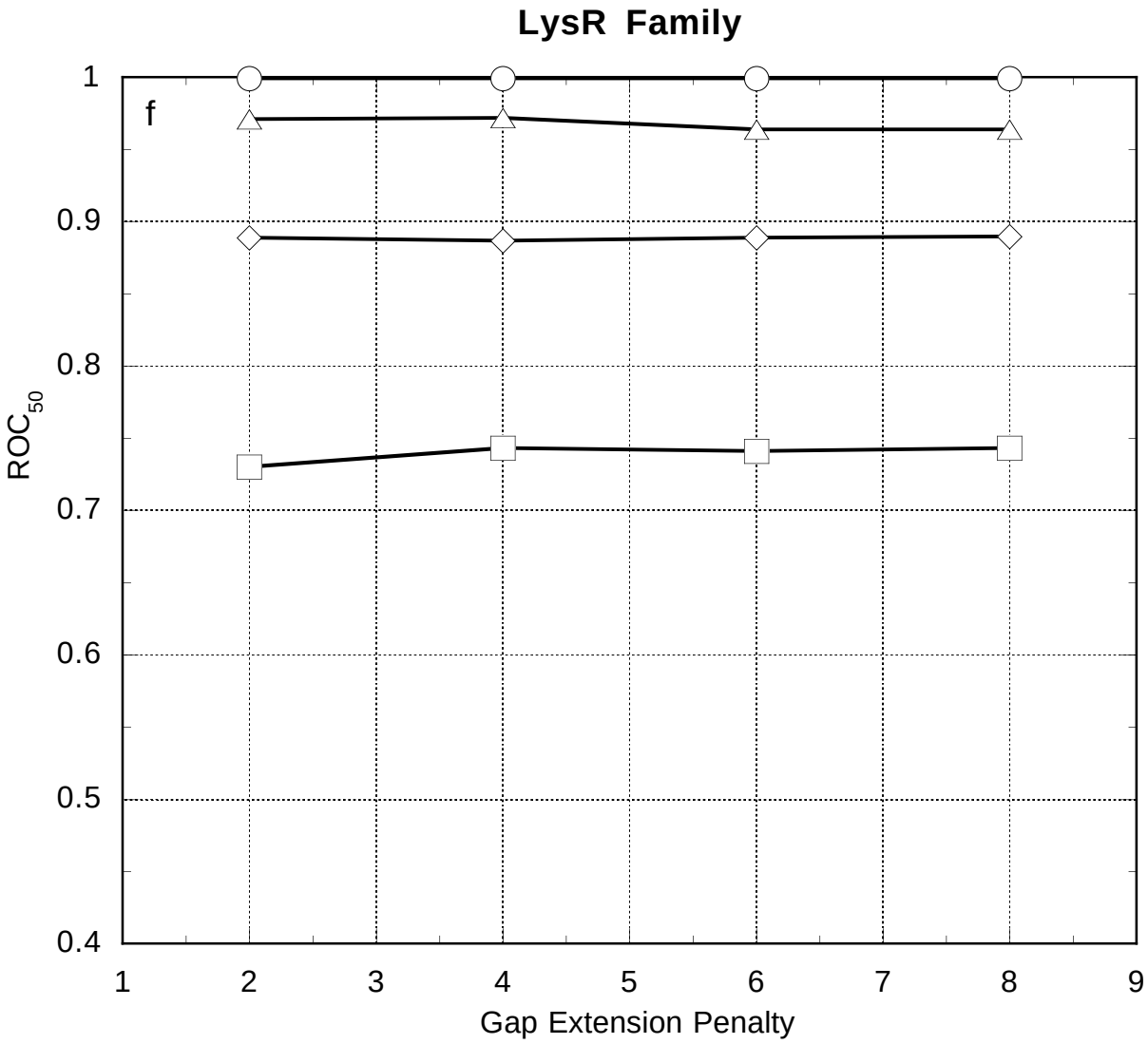

LysR Family
f
1
0.9
0.8
0.7
0.6
0.5
0.4
ROC50
1
2
3
4
5
6
7
8
9
Gap Extension Penalty

Fig 2g

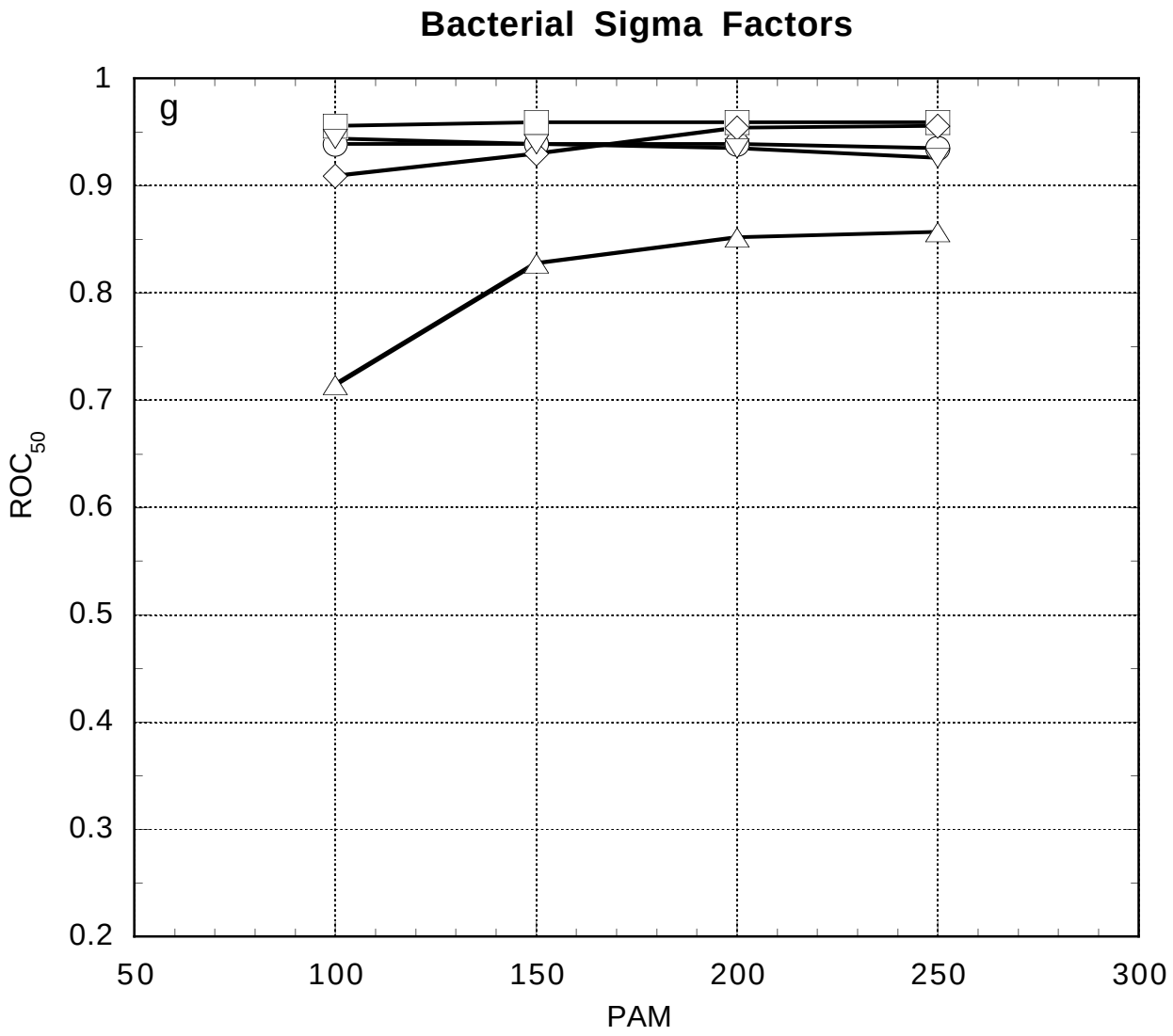

Bacterial Sigma Factors
g
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
ROC50
50
100
150
200
250
300
PAM

Fig 2h

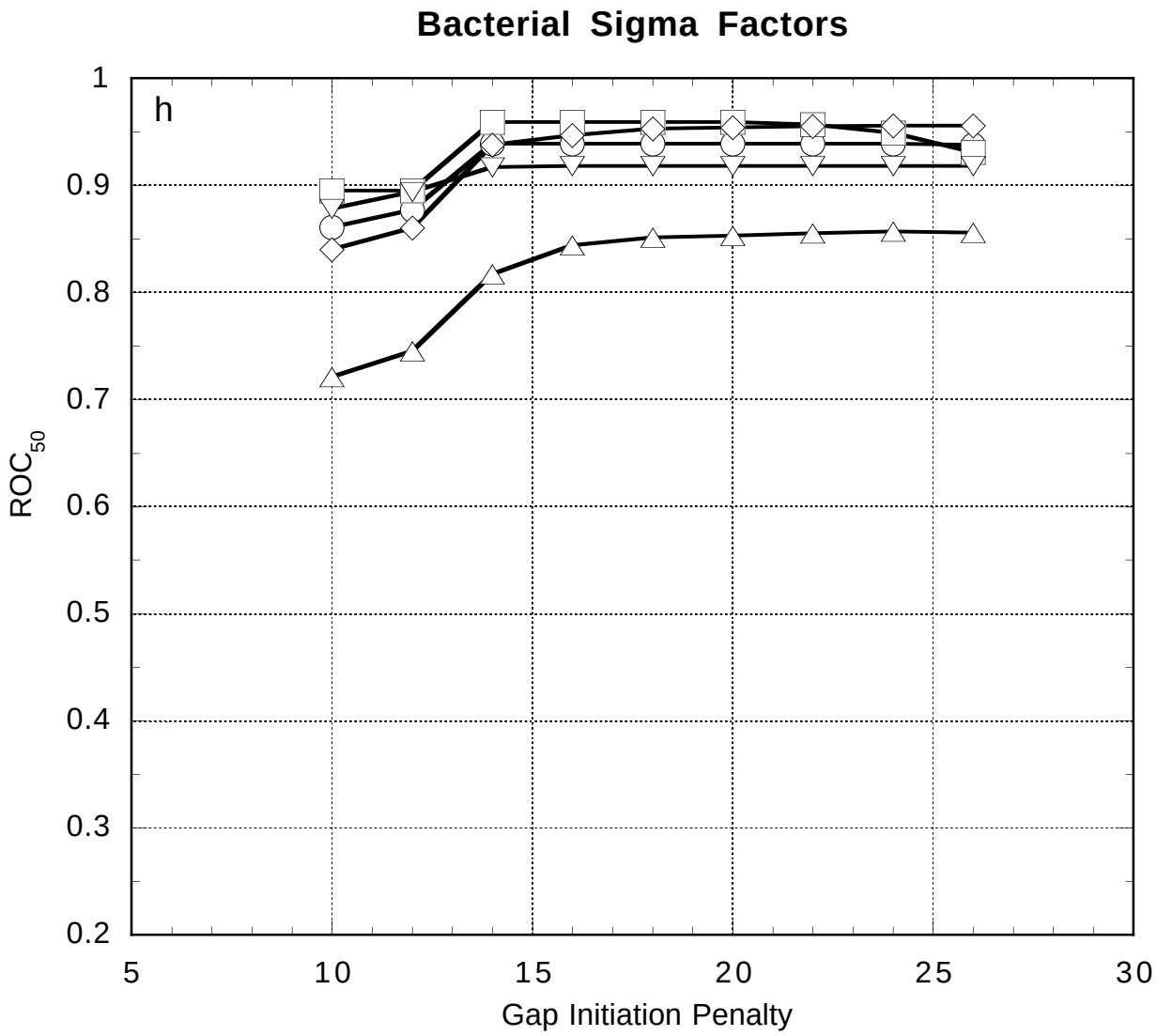
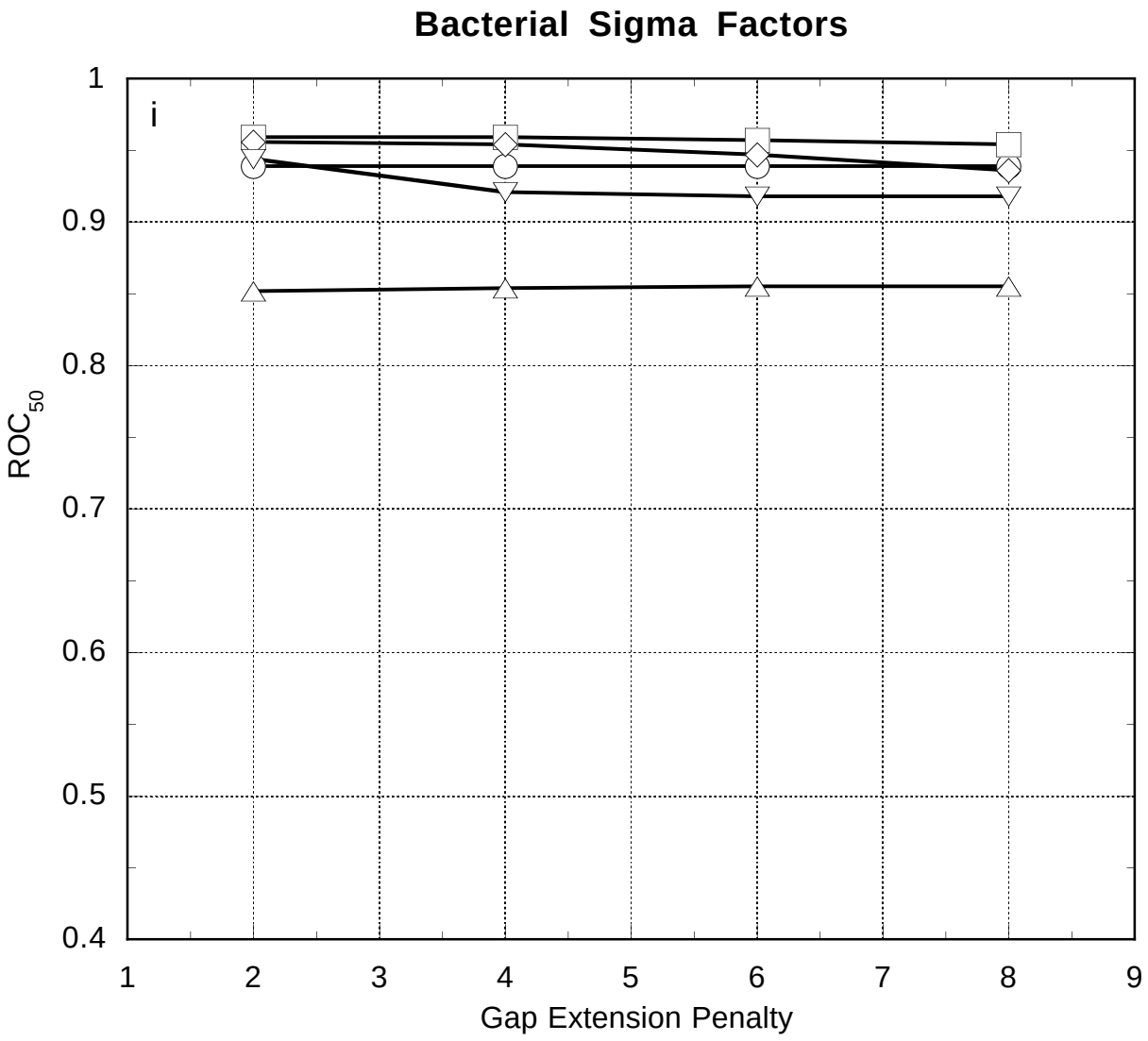

Bacterial Sigma Factors
h
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
ROC50
5
10
15
20
25
30
Gap Initiation Penalty

Fig 2i


Bacterial Sigma Factors
i
ROC50
1
0.9
0.8
0.7
0.6
0.5
0.4
1
2
3
4
5
6
7
8
9
Gap Extension Penalty