Extremophiles (2007) 11:719–732
DOI 10.1007/s00792-007-0093-y

ORIGINAL PAPER

Molecular characterization of cold adaptation based on ortholog protein sequences from *Vibrionaceae* species

Steinar Thorvaldsen · Erik Hjerde · Chris Fenton · Nils P. Willassen

Received: 14 November 2006 / Accepted: 15 May 2007 / Published online: 19 June 2007

Abstract A set of 298 protein families from psychrophilic *Vibrio salmonicida* was compiled to identify genotypic characteristics that discern it from orthologous sequences from the mesophilic Vibrio/Photobacterium branch of the gamma-Proteobacteria (*Vibrionaceae* family). In our comparative exploration we employed alignment based bioinformatical and statistical methods. Interesting information was found in the substitution matrices, and the pattern of asymmetries in the amino acid substitution process. Together with the compositional difference, they identified the amino acids Ile, Asn, Ala and Gln as those having the most psycrophilic involvement. Ile and Asn are enhanced whereas Gln and Ala are suppressed. The inflexible Pro residue is also suppressed in loop regions, as expected in a flexible structure. The dataset were also classified and analysed according to the predicted subcellular location, and we made an additional study of 183 intracellular and 65 membrane proteins. Our results revealed that the psychrophilic proteins have similar hydrophobic and charge contributions in the core of the protein as mesophilic proteins, while the solvent-exposed surface area is significantly more hydrophobic. In addition, the psychrophilic intracellular (but not the membrane) proteins are significantly more negatively charged at the surface. Our analysis supports the hypothesis of preference for more flexible amino acids at the molecular surface. Life in cold climate seems to be obtained through many minor structural modifications rather than certain amino acids substitutions.

Keywords Psychrophiles · Vibrio · Comparative genomics · Substitution matrix · Physicochemical properties

Communicated by A. Driessen.

S. Thorvaldsen (✉)
Department of Mathematics and Statistics,
Faculty of Science, University of Tromsø,
9037 Tromsø, Norway
e-mail: steinart@math.uit.no

E. Hjerde · C. Fenton · N. P. Willassen
Department of Molecular Biotechnology,
Faculty of Medicine, University of Tromsø,
9037 Tromsø, Norway
e-mail: erikh@fagmed.uit.no

C. Fenton
e-mail: chrisf@fagmed.uit.no

N. P. Willassen
The Norwegian Structural Biology Centre,
University of Tromsø, 9037 Tromsø, Norway
e-mail: nilspw@fagmed.uit.no

Introduction

The analysis of cold adaptation has been parsimonious because of lack of genome sequences from cold-adapted organisms, compared with the mesophilic and thermophlic cases. The first draft genome sequences of the cold-adapted Archaea *Methanogenium frigidum* and *Methanococcoides burtonii* were published in 2003 by Saunders et al. Their optimum growth temperature (T_{opt}) is 15 and 23°C, respectively. The psychrophilic genomes *Desulfotalea psychrophilia*, *Colwellia psychrerythraea*, and *Pseudoalteromonas haloplanktis* have also been published lately (Rabus et al. 2004; Methe et al. 2005; Medigue et al. 2005).

The family *Vibrionaceae* comprises bacteria inhabiting aquatic environments, especially marine and estuarine waters, where they often are linked with organisms ranging

from fish to plankton. Currently this family includes, among others, the genera *Vibrio* and *Photobacterium*. The vibrios are one of the species with the greatest amount of published genomes, reaching 5 competed genomes last year, and 13 ongoing whole genome sequencing projects including the cold-adapted *Vibrio salmonicida*. *V salmonicida* is one of the vibrios with the lowest known T_{opt}, 15°C on solid media and 10°C in liquid media (Colquhoun et al. 2002). The group of *Photobaceria* has one completed genome, the cold-adapted marine *Photobacterium profundum* (Vezzi et al. 2005; Campanaro et al. 2005), and two ongoing genome projects.

In our study, we focused on orthologous protein data from a relatively narrow range of closely related species belonging to the group *Vibrionaceae* of gamma proteobacteria—a strategy also adopted by Saunders et al. (2003) and Bae and Phillips (2004). While focusing on orthologous proteins has certain advantages, it also greatly reduces the number of sequences available for analysis. To meet this need, we take advantage of the recent whole-genome sequencing projects to compare protein sequences of psychrophiles and mesophiles, to yield sufficient sample size for statistical modeling. Protein families from five mesophilic and two psychrophilic bacterial genomes from the *Vibrionaceae* were aligned and analysed.

Although the study of the psychrophiles is still in its infancy, both physiological and structural adaptations already appear to be important for life in cold climate. The present lack of consensus among studies (D'Amico 2002) has given rise to the recognition that no sets of simple factors distinguish all mesophile and psychrophile proteins. If there are general rules to cold adaptation and functionality, a broader approach to the problem will be required to elucidate them. We apply a bioinformatical approach and make use of computational univariate statistical methods to extract the relevant information with respect to cold adaptation. Different new statistical analysis based on a range of physicochemical attributes of the amino acids has been used in this study (Thorvaldsen et al. 2006). We also studied amino acid composition and substitution from the aligned sequences and performed a comparative statistical analysis. This approach is based on comparisons to extract important features for thermal adaptation from the background or random differences. It should be noted that the statistical analysis described in this work can only detect general factors of cold adaptation and overlooks the subtle structural modifications that can be identified by detailed single-family structural comparisons.

The molecular mechanisms of cold adaptation are still relatively unknown, and each protein family may adopt different structural strategies to adapt to low temperatures (Gianese et al. 2002). However, some common trends of the enzyme families studied so far has been reported: The number of ion pair interactions, the side chain contribution to the exposed surface, and the hydrophobic fraction of buried residues are all observed to decrease with adaptation to low temperature (Gianese et al. 2001, 2002). The surface is also observed to be somewhat more hydrophobic, and in some protein families the volume of cavities is increasing. From previous studies it also appears that optimization of the surface charge distribution, with an excess of negative charges at the surface of the molecule, is one of the strategies of cold adaptation, because it will improve the interactions with the solvent, and thus the elasticity and fluctuation of the protein (Marx et al. 2004). The ease of this ability to change shape reduces the energy costs of the conformational changes required to interact with the substrate, and is important for the efficiency in the process. Therefore, the current accepted main hypothesis suggests that cold adapted enzymes have to increase their flexibility in order to compensate for the low thermal energy provided by the surroundings. Cold adapted enzymes are also likely to have more unfavourable (and thus unstable) secondary structures. Flexibility may also involve an increased number and clustering of flexible Gly residues, a decrease in rigid Pro residues, and eventually a reduction in Arg residues, capable of forming multiple electrostatic interactions (Fields 2001). However, the hypothesis of increased flexibility of cold active proteins still lacks a direct experimental verification, even if inflexible Pro has been used to increase the protein stability in the several mutational studies.

Evidently, each protein may use a few of the structural alterations mentioned above to acquire the required flexibility to be more or less adapted to the temperature of the environment.

Previous attempts to identify the amino acid substitutions preferred at different temperatures have compared a relatively small number of protein sequences from a wide variety of organisms. Argos et al. (1979) compared the sequences of three proteins from a variety of taxa living at different temperatures, Gianese et al. (2001, 2002) compared homolog structures from 7 and 21 different enzymes, and in a recent paper 60 thermophilic structures and sequences were compared with their mesophilic homolog (Sadeghi et al. 2006).

Several large-scale studies have also compared thermophile organisms with different growth temperatures to achieve a closer insight on protein thermostability at high temperatures (McDonald et al. 1999; Chakravarty and Varadarajan 2000, 2002). Some of these studies have focused on comparison within closely related lineages, mesophilic and thermophilic *Methanococcus* species (Haney et al. 1999), two mesophilic *Corynebacterium* species (*Corynebacterium efficiens* and *Corynebacterium glutamicum*) with slightly different optimum temperatures

for growth (Nishio et al. 2003), and two closely related hyperthermophilic genus from the Thermococcus order with optimum temperature for growth of 85°C and 98–103°C, respectively (Fukui et al. 2005). Such large-scale comparative analysis among closely related psychrophiles has not been reported.

Alignment-free sequence analysis has also been used in previous studies to compare amino acid compositions in whole genome and proteome datasets (Karlin et al. 2002; Pe'er et al. 2004). However, in our study we mainly employed alignment-based methods for examination of differences at the molecular level by comparing amino acid divergence in the protein sets. We selected 298 aligned protein families commonly existing among *Vibrionaceae*, and extracted specific amino acid replacements and a range of physicochemical attributes. The present work confirms the importance of some of the factors identified from earlier analyses and, in addition, identifies some new factors possibly responsible for enhanced protein activity in cold climate.

Materials and methods

Data for comparative analysis

The potential proteins from about 12% of the completed genome of *V. salmonicida* were compared against the corresponding genes of the other sequenced *Vibrionaceae* genomes, *V. cholerae* (Heidelberg et al. 2000), *V. fischeri* (Ruby et al. 2005), *V. parahaemolyticus* (Makino et al. 2003), *V. vulnificus* CMCP6 (van Passel et al. 2005) and *P. profundum* (Vezzi et al. 2005; Campanaro et al. 2005) in order to identify homologues. From the genome of *V. salmonicida*, 529 potential coding sequences (CDS) were aligned with the other sequenced *Vibrionaceae* proteins by running a Python script using Blast and T-coffee (Notredame et al. 2000). All alignments were inspected and verified manually for a minimum cut-off score of 40% identity with all other sequences. Proteins with low identity were eliminated because of uncertain alignments. No attempt was done to remove paralogs. The corresponding amino acid sequences of the *Vibrionaceaes* were extracted in 298 final alignments, each of 7 sequences.

Relationship of the species of *Vibrio* and relatives, based on the 16S rRNA gene sequence, is shown in Garrity (2005, vol 2B p 517). To avoid oversampling and statistical dependences in the data, one of the two genomes of *V. vulnificus* (YJ016) was excluded from the statistical analysis. The resulting sample was considered to be representative. The genomes were divided in three temperature classes based on T_{opt}: psychrophile (*P. profundum* and *V. salmonicida*), intermediate (*V. fischeri*) and mesophile (the rest). *V. fischeri* was placed in an intermediate class because of its psychrotolerant physiologies. Temperatures were assigned according to those of optimal growth or normal living environment for the species involved.

To look for sequences with experimentally known sub-cellular annotation, we performed a manual Swiss-Prot search for all sequences from *V. salmonicida*. In the further work we also made extensive use of sequence based predictors developed in the latest years, and the sub-cellular location of all the remaining sequences was automatically predicted using CELLO (Yu et al. 2004), with a threshold score of ≥3 for all sequences in the alignment. Although transmembrane proteins are known to be less conserved in different species, the overall prediction accuracy of CELLO is close to 89% (Yu et al. 2004). Surface and secondary structure were predicted using SABLE (Adamczak et al. 2004) with the sequence from *V. cholerae* as input. The secondary structure was predicted with default settings, and the solvent accessible surface area (ASA) was predicted with 0-buried core; 1,2-intermediate twilight region and 3,5,6,7,8,9-fully exposed surface. With these cut-off values the number of residues will be approximately equal in each of the three states, and the thresholds between the states are around 10 and 40% ASA (Pollastri et al. 2002). The algorithm for ASA prediction has accuracy of about 77% for a two state prediction, with ASA of 25% separating the buried and exposed residues (Adamczak et al. 2004). However, the algorithm avoids definition of discrete classes, enabling predictions for a range of different ASA. The algorithm of the secondary structure prediction has an accuracy of close to 78% (Adamczak et al. 2005). Using these sequence based predictors make it possible to analyse the alignment data relative to both its secondary and some of its three-dimensional (3D) structural constraints:

- cellular location (intracellular, membrane, extracellular)
- secondary structure (alpha, beta, loop)
- 3D structure location (core, twilight zone, surface)

As long as the errors made by the predictors are unbiased, a statistical approach will handle them in a proper way. We looked for patterns in the data following from this partitioning and decompositions, and made a special study of the predicted intracellular and membrane proteins from cold-adapted organisms.

Changes in amino acid composition

We examine the change in composition of amino acids in the different temperature populations. For each gene family, the mean change of amino acid frequency, Δf, was computed from the mean values in the psychrophilic population P, and the mesophilic population M, by this formula:

$$\Delta\overline{f}(x) = \overline{f}^{(P)}(x) - \overline{f}^{(M)}(x)$$

where x stands for the 20 amino acid residues. The total number of gene families is 298, each with three mesophilic and two psychrophilic sequences, and may be compared statistically as well as graphically (Fig. 2).

However, the analyses of compositions have serious limitations. They simply indicate the degree of evidence for an over- or under-representation of amino acids, and are not sufficient for answering other more detailed questions about the data. One should also study the nature and effects of these differences.

Amino acid substitution pair bias

By comparative genomics it is also possible to do a more detailed analysis and identify genome-wide amino acid directional biases in substitutions among proteins from different sources of interest. First we examine the substitutions of amino acids between pairwise aligned sequences. A *substitution pair* (SP) is defined as the ordered combination of two amino acids, (x_M, y_P), where residue x in population M is converted to residue y in population P. For a given pair of amino acids, the substitution order refers to the $M \rightarrow P$ direction. The number of the SP $x \rightarrow y$ can be computed as:

$$n_{x,y} = \sum (x, y), \quad x \neq y$$

The *SP-matrix* is the count of all such pairs observed in the alignment by summing over all sites. This accumulated array contains the occurrence of all position specific pairing of residues. When there are multiple sequence samples in a population, we use the average countings in the representative SP-matrix:

$$\overline{n}_{x,y}^{(M,P)} = \frac{1}{G_{MP}} \sum n_{x,y}, \quad x \neq y$$

where G_{MP} is the number of ordered pairs of gene sequences in the two populations. The use of sample mean also removes some of the random genetic drift (noise).

We may both calculate SP-matrices *between two* populations, as well as *within one* population. Gaps may be counted as an ''extra'' amino acid.

It is well known that amino acid substitutions between mesophilic and thermophilic organisms are not all symmetrical (Haney et al. 1999; McDonald et al. 1999). This is commonly interpreted as an indication that certain amino acids are favoured over others by adaptive selection at different temperatures, and a natural question is witch substitutions are over- or under-represented compared to a random model. The substitutions of residue $x \rightarrow y$, may be computed versus the following backgrounds (cf. Chakravarty and Varadarajan 2002):

1. (x_M,y_P) versus (y_M,x_P) [forward ($M \rightarrow P$) versus reverse ($P \rightarrow M$)]
2. (x_M,y_P) versus (x_M,y_M) [forward ($M \rightarrow P$) versus internal ($M \rightarrow M$)]
3. (x_M,y_P) versus ($x_{M'}$,$y_{P'}$) [forward ($M \rightarrow P$) versus forward ($M' \rightarrow P'$)]

M' and P' are independent of the original dataset. The probability of a directional amino acid bias greater than or equal to that observed is compared relative to several relevant background frequencies (control data). For this purpose 2×2 contingency tables were constructed, with the number of aligned sites exhibiting each of the possible pairwise substitution patterns.

Since there is more than one sequence in each population, we combine all the replicated data by using the mean cell counts in a representative 2×2 table. For purposes of illustration we consider model 2 above. The statistical test is computed from Table 1.

P-values were calculated by the two-tailed Chi-square test. In model 1 above, only two sequences are needed as input for an independent cell counting, and the biased mutations from mesophiles to psychrophiles are extracted by simply analysing the differences between the SP-matrix and its transposed matrix. In model 2 we need minimum three sequences as input, and we have to use two separate SP-matrices. Testing of multiple SPs will potentially increase the number of false positives (Type I error), but similar results from models 1 and 2 will add weight to the conclusions. The statistical power of these tests depends strongly on the sample size and the composition of the sample (Fleiss et al. 2003), and instead of comparing two tests with low power, model 3 was used to find dissimilar adaptations in membrane proteins relative to the intracellular proteins.

Amino acid properties

The amino acids have many different physicochemical characteristics, often denoted by propensities or properties, and the amino acid alphabet can be efficiently reduced based on the physicochemical properties of each acid. All property scales assigns specific numerical values to each of the 20 amino acids (e.g., its volume or hydrophibicity). An alignment is a set of matched *sites* where there is a meaningful one-to-one correspondence between the data points in one sequence and those in the others. This gives us the possibility to investigate the property differences in the sequences by statistical methods. A non-parametric cumulative Mann–Kendall trend test was carried out for systematic comparison of various traits of the sequences

Table 1 Example of how the Chi-square test (with 1 *df*) for two independent proportions is carried out for the comparison of population mean evaluated on a mesophilic background

Substitutions of x	y	Not y
Adaptation $M \rightarrow P$	$\overline{n}_{x,y}^{(M,P)}$	$\overline{n}_{x,\text{non}-y}^{(M,P)}$
Background $M \rightarrow M$	$\overline{n}_{x,y}^{(M,M)}$	$\overline{n}_{x,\text{non}-y}^{(M,M)}$

This testing scheme removes some of the phylogenetic noise that may exist in the data

between the groups of ortholog proteins in this study, as described in an earlier paper (Thorvaldsen et al. 2006).

A comprehensive set of properties should cover many of the functional and structural aspects of proteins. The most relevant propensities related to secondary structure are alpha helical tendencies, beta sheet tendencies, and coil tendencies (Chou and Fasman 1978). Furthermore, the amino acids were divided in four main groups: negative charged, positive charged, polar and hydrophobic. For the property *negative charge* Asp and Glu were assigned a charge magnitude of 1, and all other amino acids were assigned a negative charge of 0. For the property *positive charge* Arg and Lys were assigned a charge magnitude of 1, His was assigned a charge of 0.3, and all other amino acids were assigned a positive charge of 0. The amino acids Ala, Val, Phe, Pro, Met, Ile and Leu were classified as hydrophobic, with Trp assigned a hydrophobic magnitude of 0.3. The rest of the amino acids were classified as polar, with magnitude 0.7 for His and Trp. For properties related to 3D structure we also used these properties: isoelectric point and polarity (Zimmerman 1968), hydrophobicity (Kyte and Doolittle 1982), molecular weight (Fasman 1976), denaturated ASA and Gibbs free energy change of hydration for native protein (Oobatake and Ooi 1993), bulkiness (Zimmerman 1968), shape as position of branch point in side-chain (Gunsteren and Mark 1992) and flexibility.

The bulkiness of an amino acid is defined as the ratio of the side-chain volume to length, which provides a measure of average cross-section of the amino acid, thus having a relevance to packing considerations. Flexibility is complex to predict. We used the new scale of peptide flexibility published by Huang and Nau (2003, 2005), which is based on photochemical techniques for measuring collisions in polypeptides, and thus should provide a relevant biochemical measure for flexibility of the backbone in the protein. Hydrophobicity of membrane proteins was calculated by the combined membrane scale of Ponnuswamy and Gromiha (1993).

Results

From the potential coding sequences of *V. salmonicida*, 298 alignments were made, which became the basic data for our investigations. The final alignments comprised of 105,586 multiple aligned amino acid sites. Of the 298 alignments, 183 were determined to be intracellular proteins, and 65 membrane proteins.

Amino acid sequence comparisons

Sequence alignments of the *Vibrio* proteins reveal a high degree of identity between the different species (Table 2, Fig. 1). The mean identity is highest at the predicted core (80–94%) and lowest at the surface (56–83%). We applied the methods of comparative analysis of protein sequences as described in the methodological part.

Amino acid composition

To examine if there were detectable trends in the amino acid composition with growth temperature, amino acid compositions of mesophilic and psychrophilic genomes were compared. Despite the high level of sequence conservation, there are some differences in composition that may be symptomatic for cold adaptation (Fig. 1). Some systematic trends are observed in the direction Meso-

Table 2 Overall sequence identity and mean sequence length of the organisms in this order from left to right: *V. vulnificus* YJ016, *V. vulnificus* CMCP6, *V. parahaemolyticus*, *V. cholerae*, *V. fischeri*, *V. salmonicidia* and *P. profundum*

	V. vul Y	*V. vul* C	*V. par*	*V. chol*	*V. fis*	*V. sal*	*P. pro*	Mean sequence length	T_{opt}(°C)
V. vul Y	1	0.975	0.855	0.807	0.738	0.728	0.698	346.0	37
V. vul C	0.975	1	0.851	0.803	0.736	0.725	0.694	340.1	37
V. par	0.855	0.851	1	0.808	0.750	0.740	0.708	343.7	37
V. chol	0.807	0.803	0.808	1	0.730	0.722	0.694	344.4	37
V. fis	0.738	0.736	0.750	0.730	1	0.893	0.716	341.6	28
V. sal	0.728	0.725	0.740	0.722	0.893	1	0.713	341.3	15
P. pro	0.698	0.694	0.708	0.694	0.716	0.713	1	344.8	15

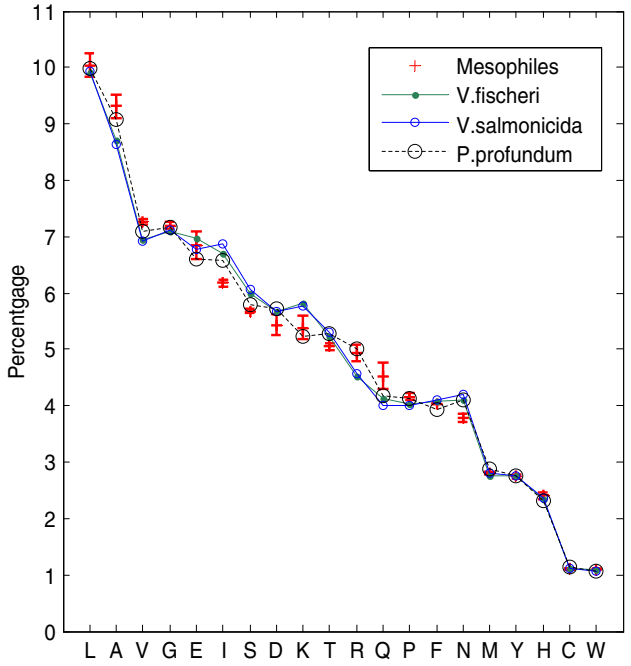

Fig. 1 Comparison of amino acid compositions, with the amino acids ranked according to frequencies in the mesophilic population. *Error bars* represent the empirical standard deviations of the mesophilic sequences. Three individual species are also shown. Amino acid abbreviations: *A* alanine, *C* cysteine, *D* aspartic acid, *E* glutamic acid, *F* phenylalanine, *G* glycine, *H* histidine, *I* isoleucine, *K* lysine, *L* leucine, *M* methionine, *N* asparagine, *P* proline, *Q* glutamine, *R* arginine, *S* serine, *T* threonine, *V* valine, *W* tryptophan, *Y* tyrosine

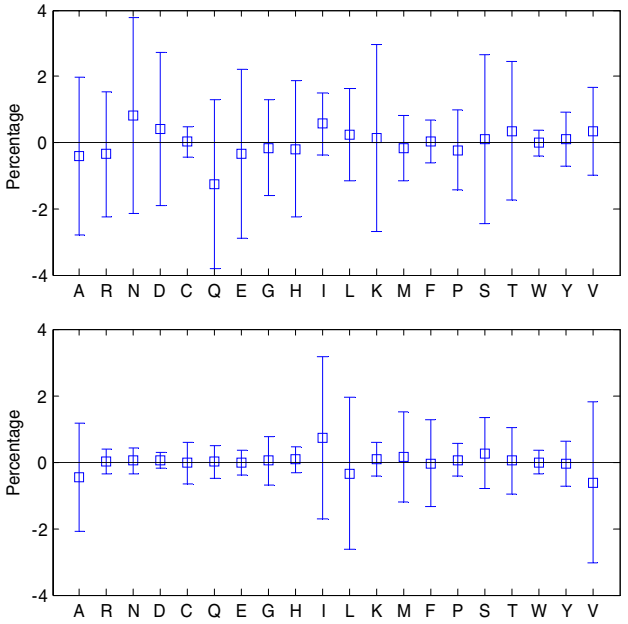

Fig. 2 Comparison of the mean compositional variations in the populations of 298 ortholog proteins, observed in the direction mesophiles → psychrophiles. The surface region of the molecules is shown at the *top*, and at the core region at the *bottom*. *Error bars* represent the empirical standard deviations

phile → Psychrophile, as measured from the mesophilic mean value. The strongest increase was for Ile (up by 0.41–0.69% points), with weaker contributions from Asn (up by 0.32–0.42% points), while Ala, Val and Gln all show corresponding decrease (down by 0.25–0.70, 0.18–0.35 and 0.36–0.52% points, respectively). The charged residues Lys and Arg appear to have a different behaviour in different species at the compositional level.

At the predicted surface, the same tendencies are observed for Ile, Asn and Gln (up by 0.43–0.60, up by 0.61–0.96 and down by 1.03–1.44% points, respectively), but Ala shows no systematic behaviour and Val is instead increasing somewhat (data not shown). In both psychrophilic and mesophilic organisms, the relative abundance in charged amino acids at the surface of the proteins are high, but not notably different.

In the present dataset, the compositional changes on the basis of each protein family were also computed. We observed that the nature of protein compositional variations differ at buried and exposed locations (Fig. 2). The overall compositional difference between the mesophilic and psychrophilic population is largely due to differences in the surface-exposed regions of proteins, in agreement with general knowledge about proteins. These results are also in agreement with mutational studies (Martin et al. 2001) that highlight the important role of surface residues in protein thermal stability. Changes in densely packed protein cores (Richards 1977) often create packing defects and are thus destabilising (Lim and Sauer 1991). Residues at the surface tend to be more flexible because of fewer intra-protein interactions, and thus their contribution to stability are often locally confined and additive.

We also noticed a stronger decrease in Pro content in the predicted loop regions (down by 0.26% points) than in the other parts of the molecule.

The compositional changes of the membrane and the intracellular proteins were also studied (data not shown). Both groups follow the same main trends as reported above, although the membrane proteins shows no increase in mean of Ser, and the intracellular proteins show no mean increase of Lys. In loop regions the mean number of Ser even decreases in the membrane proteins.

Specific amino acid substitutions

The preceding compositional analysis overlooks the importance of amino acids that are gained in some contexts and lost in others. To better understand the individual contributions by substitutions, Fig. 3 reports all 380 SPs (421 when gaps are included). The SPs were all tested for statistical significance in both models 1 and 2 by using the

two-sided Chi-square test, as defined in section ''Materials and methods'' above, and ranked according to their *P*-value in model 2. In Table 3 we present the most significant substitution biases in the ($M \rightarrow P$) direction. The tests were performed with an assumption that the significant changes seen in substitution patterns may be interpreted as due to thermal adaptation.

In terms of involvement in significant pairwise substitutions, Ile, Glu, and Ala are the most central residues (Fig. 3). Ala is repeatedly substituted by a variety of other residues consistent with the reduced Ala content in psychrophiles. Ala in mesophilic *Vibrio* proteins is especially replaced by Ile, but avoids Gln and Glu in the psychrophilic counterparts, with noteworthy directional bias. Substitution bias is more prominent at exposed regions than a buried site. At exposed sites, the approximately isosteric substitution Ser $\rightarrow$ Ala are avoided, and Ser are instead preferentially substituted by Ile, with increased side chain. Most prominent substitutions at exposed sites, and in the overall molecules, involve charged residues (Asp, Glu, Lys) and are preferentially substituted with hydrophobic Ile, polar Asn or positive charged Lys. This is consistent with the increased content of Ile and Asn in psychrophiles. The positive residues Lys and Arg behave quite differently with respect to substitution. Arg is rarely substituted, whereas Lys is preferentially substituted with Ile. Most of the significant enhanced or suppressed substitutions are observed both at the surface and in the entire data set. In the loop regions we found that substitutions of Ala avoids inflexible residue Pro (*P*-value = 0.044 in model 2, and *P*-value = 0.012 in model 1).

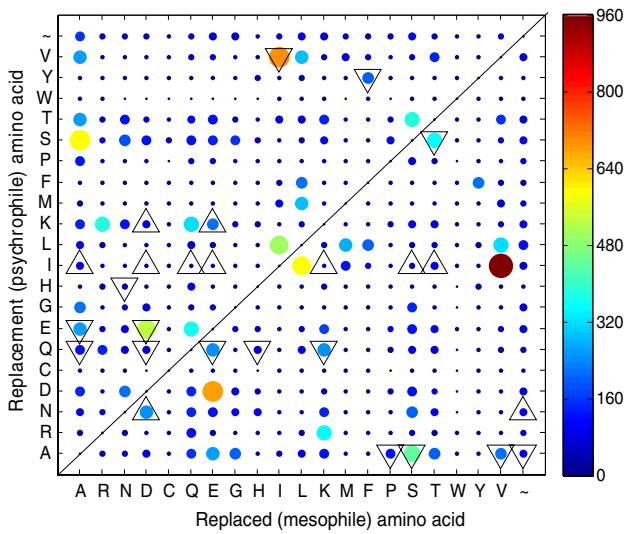

Fig. 3 Visualisation of the number of all the pairwise substitutions observed in the comparison of 298 ortholog proteins between the mesophilic and psycrophilic groups. The *size* and *colour of each marker* indicates the magnitude of the substitution (see colour-bar). Favoured substitutions (*P*-value < 0.05) in the mesophile $\rightarrow$ psychrophile direction (described as model 2 in the text) are also marked with *upward-pointing triangles*, and non-favoured substitutions (*P*-value < 0.05) with *downward-pointing triangles*. Amino acid abbreviations are as in Fig. 1, and a *tilde* (~) indicates deletion

Table 3 also summarises the most frequently observed replacements in number. It appears that most of the substitutions are for hydrophobic amino acids between themselves, or for non-hydrophobic amino acids between themselves. The most frequent substitutions are between: Val–Ile, Glu–Asp, Ala–Ser and Leu–Ile. However, many of the often occurring replacements are also commonly observed among mesophilic *Vibrionaceae*, indicating that they do not correlate with a typical procedure of cold adaptation. The observation suggests frequent substitutions of these amino acids between *Vibrionaceae*.

The ratio of forward substitutions to reverse substitution for a given pair of amino acids, relative to its background distribution, reveals the most biased exchanges. These biased substitutions suggest distinct selective pressures on the amino acids between the mesophilic and psychrophilic population. However, most of these SPs are so rare, even in the favoured direction, that they cannot be the major contributor to protein thermal adaptation. Only Ile–Val contribute an average of nearly one net replacement per typical 300-aa protein, and only 16 other biased replacements contribute a net compositional shift of 30 or more residues in the entire data set (one per second or third typical protein).

To find a better statistical approach to detect substitution biases, it appears natural from Fig. 3 to pool the substitution data that share the same outcome in the psychrophilic population, e.g. to study the substitutions $x \rightarrow$ aa, where x may be any amino acid. The results given in Table 4 show with strong significance (*P*-value < 10^{-6}) that Ile, Asn, Ala and Gln has the most psycrophilic involvement. Ile and Asn are enhanced whereas Gln and Ala are suppressed. We also notice that Pro is significantly suppressed in loop regions (*P*-value = 0.04 in model 2, and *P*-value < 10^{-6} in model 1).

The difference in substitution pattern in the membrane and the intracellular proteins were also compared. A significant difference was noted for the Ser $\rightarrow$ Ala substitution (*P*-value = 0.02), witch is avoided in intracellular proteins and not in membranes, and for Ala $\rightarrow$ Glu (*P*-value = 0.05), that is avoided in membranes and more preferred in intracellular proteins.

Physicochemical properties

We also investigated further how strongly some of the physical, chemical, and geometric properties of the 20 amino acids can be attributed to thermal adaptation. The alignments were analysed for features that have been pro-

Table 3 Amino acid replacements in the dataset of 298 protein alignments that are most biased in difference between mesophile → psychrophile and mesophile–mesophile (model 2), with the comparison based on psychrophile → mesophile (model 1) shown in parentheses

Most biased substitutions				Most frequent substitutions in number			
Pair	Forward	Reverse	*P*-value model 2 (1)	Pair	Forward	Reverse	*P*-value
EQ	252	363	$<10^{-6}$ ($<10^{-6}$)	VI	962	696	0.35
DE	536	689	0.00001 (0.031)	IV	696	962	0.0013
EK	213	171	0.00003 (0.14)	ED	689	536	0.58
KQ	249	322	0.00004 (0.013)	AS	599	441	0.39
AI	93	59	0.00060 (0.096)	LI	589	499	0.55
EI	40	23	0.00064 (0.059)	**DE**	536	689	0.00001
SA	441	599	0.00077 (0.00039)	IL	499	589	0.98
AQ	134	164	0.00099 (0.00072)	**SA**	441	599	0.00077
HQ	85	107	0.0012 (0.013)	ST	390	351	0.97
IV	696	962	0.0013 (0.19)	RK	379	359	0.19
TI	103	78	0.0016 (0.0042)	QE	363	252	0.061
AE	268	269	0.0097 (0.024)	KR	359	379	0.081
DQ	77	144	0.010 (0.00049)	TS	351	390	0.021
~A	95	154	0.012 (0.00002)	VL	331	295	0.64
SI	44	32	0.017 (0.062)	QK	322	249	0.89
~N	109	64	0.018 (0.0016)	LM	296	281	0.29
FY	199	212	0.018 (0.34)	LV	295	331	0.10
TS	351	390	0.021 (0.65)	ML	281	296	0.087
DK	80	71	0.02 (0.087)	EA	269	268	0.096
QI	39	22	0.035 (0.19)	**AE**	268	269	0.0097
PA	132	128	0.036 (0.38)	AT	265	199	0.84
VA	216	256	0.039 (0.00077)	AV	256	216	0.70
DI	13	9	0.040 (0.19)	**EQ**	252	363	$<10^{-6}$
DN	251	212	0.041 (0.0002)	**DN**	251	212	0.041
KI	41	27	0.042 (0.048)	**KQ**	249	322	0.00004
NH	65	113	0.048 (0.050)	LF	219	204	0.70

The right part of the table shows the substitutions that are most frequent in numbers, with *P* values from model 2. Substitutions that are found to be significant (*P* value < 0.05) in both models are shown in bold. A tilde (~) indicates deletion

posed as important for protein function at low temperature (Georlette et al. 2004; Marx et al. 2004). In the examination of how different physicochemical properties may contribute to cold adaptation, we used 14 different properties.

Tables 5, 6, and 7 list the strength in the statistical bias between the temperature populations of the 183 intracellular and the 65 membrane protein families, for all properties and in different regions. *P*-values were calculated by a two-tailed cumulative Mann–Kendall trend test. When analysing all the 298 proteins in one pool, the results essentially came out as for the intracellular case (Table 6), and are not shown.

Significant differences are located both in the secondary regions, and in the core, twilight and surface of the molecule. Our calculations on the amino acids forming helices and beta-strands indicate a lower degree of stabilisation for psychrophiles compared with mesophiles both in intracellular and membrane proteins (Table 5). The surface hydrophobicity is also increased significantly in both psychrophilic protein collections, but the isoelectric point property acts differently.

Discussion

We performed comparative analysis using protein alignments from the *Vibrionaceae* to find general characteristics for a cold-adapted organism. The use of proteins from the same taxonomy reduced problems associated with physiology and genetic diversity that have been a problem in other studies, and the model is robust against differences due to differences between different phylogenetic lineages. Eventual horizontal gene transfer will make this more complicated, but environmental adaptations of *Vibrionaceae* species are suspected to be driven mainly by vertical

Table 4 *P*-values from pooled substitutions, where the amino acids are the outcome in the psychrophilic population.

Subst.to aa	Model 1		Model 2	
	Changes in entire dataset	Regions with *P*-value < 0.05	Changes in entire dataset	Regions with *P*-value < 0.05
A	↓ ($<10^{-6}$)	a,b,l,c,t,s	↓ (3×10^{-6})	a,l,t,s
R	↓ (0.002)	a,l,s	↓ (0.04)	l
N	↑ ($<10^{-6}$)	a,l,t,s	↑ (5×10^{-4})	a,l,s
D	↑ (5×10^{-5})	a,l,c,t	↑ (0.29)	t
C	↑ (0.23)		↑ (0.08)	
Q	↓ ($<10^{-6}$)	a,b,l,s	↓ ($<10^{-6}$)	a,b,l,s
E	↓ (0.01)	a,s	↓ (1×10^{-4})	a,s
G	↓ (0.19)	l,t,s	↑ (0.08)	
H	↓ (0.05)	l,t,s	↓ (0.06)	l
I	↑ ($<10^{-6}$)	a,b,l,c,t,s	↑ ($<10^{-6}$)	a,b,l,s
L	↓ (0.18)	a,c,t, ↑(s)	↑ (0.15)	s
K	↑ (0.04)	c	↑ (0.02)	t,s
M	↑ (0.38)	c, ↓(s)	↑ (0.04)	
F	↓ (0.71)		↑ (0.02)	
P	↓ (0.01)	l,s	↓ (0.41)	l
S	↑ (0.0001)	a,b,c,t	↓ (0.72)	l
T	↑ (2×10^{-4})	l,t	↑ (0.06)	
W	↓ (0.14)	b	↑ (0.30)	
Y	↑ (0.67)		↑ (0.02)	l
V	↓ ($<10^{-6}$)	a,b,l,c,t, ↑(s)	↓ (0.57)	c,t, ↑ (s)
~	↓ (0.15)	↑ (a,b,l,c,t,s)	↓ (0.63)	l, ↑(s)

Substitutions with *P*-value < 0.05 in both models are shown in bold. Particular secondary and 3D regions in the hosting molecule are also marked if they have specific regional significance: a = alpha helix, b = beta sheet, l = loop, c = core, t = twilight zone and s = surface. The tilde (~) in the last line indicates deletion, but deletions mostly fall outside regions with known structure

Table 5 Changes in secondary structure propensity and peptide flexibility from mesophile to psychrophile populations of intracellular and membrane proteins, with *P*-values from the cumulative trend test in parenthesis

Region	Alpha	Beta	Loop
Propensity intracellular	**↓** ($<10^{-6}$)	**↓** (0.0002)	↓ (0.25)
Propensity membrane	↓ (0.004)	↓ (0.05)	↑ (0.93)
Flexibility intracellular	↑ (0.45)	↓ (0.42)	↑ (0.08)
Flexibility membrane	↓ (0.003)	↑ (0.58)	↓ (0.91)

Regions that are highly significant (*P*-value < 0.001) are marked with arrow in bold (i.e. **↓**)

processes like substitutions, rather than horizontal gene transfer (Thompson et al. 2004). And instead of focusing on the details of a specific subset of proteins, we calculated mean values for the total dataset in order to determine whether general trends were apparent.

There are both similarities and differences between our finding and those reported previously. One of the main differences is an increase in the Ile and a decrease in the Gln amino acid content. However, large variation in Gln is also observed in the mesophilic dataset, and this may possibly eliminate its role as a mean to enhance the protein low-temperature adaptation.

Properties of the most frequent SPs

Above we presented a statistical analysis of the most biased amino acid substitutions pairs, and many of these SPs consist of the most hydrophilic/hydrophobic residues, like (Asp, Gln, Glu, Lys → Ile). In general, residues forming the most biased SPs have extreme values of one or more essential physicochemical properties.

In terms of involvement in significant SPs, Ile is the most mutable residue, being involved in 8 of 26 pairs of Fig. 3. The psychrophile population contains a larger number of Ile, even in the core of the molecule. If Ile and Leu residues are compared, these two residues have the highest (and equivalent) partial specific volumes of the hydrophobic residues. In proteins, the Leu side chain is most often found in two of its three rotamer conformations (χ_1 of 180° or 300°). The Ile side chain frequently adopts four different rotamer conformations, and the three χ_1 values are found. With this conformational flexibility, Ile might be better able to fill various cavities that can occur during protein core packing (Britton et al. 1995; Lovell et al. 2000; Kazuoka et al. 2003). Similar considerations can be made for the Val and Thr residue, which mainly adopts only two of their three different rotamer conformations.

Table 6 Changes in number and physicochemical values of intracellular proteins in various molecular layers from mesophile to psychrophile populations

Property	Core	Twilight	Surface
No. of positive charge aa	↑ (0.001)	↑ (0.86)	↓ (0.06)
No. of negative charge aa	↑ (0.01)	↑ ($<10^{-6}$)	↑ (0.02)
No. of charged aa	↑ ($<10^{-6}$)	↑ (0.002)	↓ (0.96)
Isoelectric point scale	↑ (0.24)	↓ ($<10^{-6}$)	↓ ($<10^{-6}$)
No. of polar aa	↑ (10^{-6})	↑ ($<10^{-6}$)	↑ (0.50)
Polarity scale	↑ (0.09)	↑ (0.40)	↓ ($<10^{-6}$)
No. of hydrophobic aa	↓ ($<10^{-6}$)	↓ ($<10^{-6}$)	↓ (0.40)
Hydrophobicity scale	↓ (0.34)	↓ (0.03)	↑ ($<10^{-6}$)
Molecular weight	↑ ($<10^{-6}$)	↑ (6×10^{-4})	↓ (0.002)
Accessible surface area	↑ ($<10^{-6}$)	↑ (0.53)	↓ (0.002)
Bulkiness	↓ ($<10^{-6}$)	↓ ($<10^{-6}$)	↑ (0.72)
Shape (position of branch point)	↓ (0.17)	↑ (0.38)	↓ (1×10^{-5})
Gibbs free energy change	↓ (0.12)	↓ (0.24)	↑ ($<10^{-6}$)
Flexibility, peptide	↓ ($<10^{-6}$)	↑ (0.75)	↑ (4×10^{-6})

P-values from the cumulative test are included in parenthesis. Regions that are highly significant (*P*-value < 0.001) are marked with arrow in bold (i.e. ↓)

Table 7 Changes in number and physicochemical values of membrane proteins in various molecular layers from mesophile to psychrophile populations

Property	Core	Twilight	Surface
No. of positive charge aa	↑ (0.11)	↑ (0.07)	↓ (0.11)
No. of negative charge aa	↑ (0.27)	↑ (0.31)	↓ (0.14)
No. of charged aa	↑ (0.02)	↑ (0.18)	↓ (0.005)
Isoelectric point scale	↑ (0.65)	↓ (0.06)	↓ (0.55)
No. of polar aa	↑ (0.13)	↑ (2×10^{-5})	↓ (0.27)
Polarity scale	↑ (0.56)	↑ (0.16)	↓ (2×10^{-5})
No. of hydrophobic aa	↓ (0.01)	↓ ($<10^{-6}$)	↑ (3×10^{-5})
Hydrophobicity scale	↓ (0.82)	↓ (0.04)	↑ (9×10^{-6})
Hydrophobicity membrane	↓ (0.46)	↓ (0.004)	↑ (0.002)
Molecular weight	↑ (9×10^{-5})	↑ (0.02)	↓ (0.06)
Accessible surface area	↑ (2×10^{-6})	↑ (0.18)	↓ (0.86)
Bulkiness	↓ (0.0002)	↓ (0.02)	↑ (0.06)
Shape (position of branch point)	↓ (0.03)	↓ (0.27)	↓ (1×10^{-5})
Gibbs free energy change	↑ (0.23)	↓ (0.90)	↑ ($<10^{-6}$)
Flexibility, peptide	↓ ($<10^{-6}$)	↓ (0.55)	↑ (0.04)

P-values from the cumulative test are included in parenthesis. Regions that are highly significant (*P*-value < 0.001) are marked with arrow in bold (i.e. ↓)

There is also a general trend for more Asn residues in psychrophiles, and this may be attributed to this residue having a great propensity at high temperatures to make proteins age by oxidative damage. Asn residues often undergo deamidation cyclisation, a process extremely sensitive to temperature (Daniel et al. 1996). An overall enrichment of Asn residue was also found in a resent study of the psychrophile *Pseudoalteromona* proteome (Medigue et al. 2005). On the other hand, the closely related Gln shows the opposite trend in occurrence, and is also able to deamidate, but at a much lower rate than for Asn (Daniel et al. 1996).

The high mutability of Ala is probably due to its ''default residue'' role. The lack of a gamma-carbon, besides contributing to the high alpha-helical propensity, also allows substitutions with small steric hindrances. Ser frequently substitute with Ala, although somewhat underrepresented. Both residues have the lowest free energy of hydration (Levitt 1976). Ser can weaken hydrophobic interaction inside proteins, because Ala is more hydrophobic in the environment of a protein than Ser (Taylor 1986). The Ser residue tends to impair hydrophobic interaction between beta-strands, while the Ala can be effective in bridging strands (Nishio et al. 2003). A lower Ala content in psychrophilic proteins can be supposed to reflect the fact that Ala is the best helix-forming residue. Ser also frequently replaces with Thr due to their enhanced capacities to form hydrogen bonds and beta-sheets and to their high apparent partial specific volumes.

To simultaneously emphasize the magnitude of substitution bias and the frequency of substitution, we examined the amino acid pairs with the largest numerical difference between forward and reverse substitutions (Table 3, right part). We may expect that these substitutions may have the broadest roles in cold adaptation. The list is dominated by conservative substitutions within its main category (charged, uncharged polar, or nonpolar), some less conservative changes are interspersed. Their frequent occurrence means that these substitutions can be accepted in many contexts, while their significant bias suggests that they are useful to cold adaptation. Yet, even these numerically most biased substitutions are far from universal. Individually, only the substitutions Val → Ile are sufficiently common to contribute a net shift of almost one residue per typical protein, and 25 other substitutions are statistically significant. Overall, the 26 SPs with *P*-values < 0.05 in Fig. 3 contribute a net shift in the forward, cold adapting direction of 1,342 residues in the 298 sequences analysed. Under the (overly simple) model that cold adaptation is caused by this alter of expected changes; these amino acid substitutions would contribute four to five changes per typical protein.

Whereas Table 3 and Fig. 3 addresses specific amino acid substitutions, we also looked for basic themes associated with the events, by the list of amino acid properties in Tables 5, 6, and 7. Interpreting the results from the property analyses are complicated by pairwise correlations

among some of the properties, ranging from 0.93 (molecular weight and ASA) to –0.77 (Shape and Gibbs free energy change). But several properties were also significantly correlated with cold adaptation; and most of the strongest correlations could be placed into some general categories, which show good agreement with other studies.

Surface and core

At the molecular surface of the intracellular molecules we find no great differences in the number of charged, polar or hydrophobic residues. But the isoelectric point is decreased very significantly. A more negative charged surface is considered beneficial for cold adaptation, because it will increase the solubility of the protein in water. In addition, the numbers of charged amino acids are found to increases significantly in the twilight region of the protein. Polarisability is a fundamental concern in solvated systems, and the surface commonly redistributes electrons in response to the surrounding electric fields.

A similar decrease of isoelectric point is not observed in the membrane proteins (Table 7). The cell membrane of gram-negative bacteria is composed of two lipid polar bilayers, which interacts with the transmembrane protein. All membranes also have a substantial fraction of negatively charged lipids (Wieslander and Rosén 2002), yielding a certain surface charge density and potential of the lipid bilayer. Since electrostatic interactions do not significantly depend on temperature, it is reasonable that the electrostatics of the membrane protein is more conserved than for cytoplasmic proteins.

It is well known that the hydrophobicity of amino acid residues plays an important role in protein folding. In aqueous environments the hydrophobic effect destabilizes unfolded forms, and the temperature corresponding to the maximum stability of a protein increase with increasing hydrophobicity of the folded core (Creighton 1991). However, it should be noted that low temperatures weaken hydrophobic effects. By comparing the surface hydrophobisity in the population of all mesophilic *Vibrionaceae* with the corresponding populations of proteins active at lower temperatures, we find that the surface is significantly more hydrophobic both in the intracellular proteins ($P < 10^{-6}$) and in membrane proteins ($P = 9 \times 10^{-6}$).

The lipid layer, with a strongly hydrophobic interior, is the structural base for all biological membranes and the surrounding environment for most transmembrane proteins. Since there are different ways of measuring hydrophobisity of amino acids, there exists an alternative combined hydrophobicity scale relative to membrane surroundings (Ponnuswamy and Gromiha 1993). To control the result above, we performed the same test by this scale, and found that this confirmed the previous result (Table 7).

The enhanced surface hydrophobicity may be compensatory for its separate fading by low temperature, or it may destabilise the entire structure and increase its disorder and flexibility. Most of the gain in hydrophobicity is caused by conservative amino acid replacements: minor changes that could increase van der Waals contacts and packing density without requiring major structural rearrangements. However, significant hydrophobicity is also contributed by substitutions of charged and polar residues in the mesophile proteins with the Ile residue in the psychrophile counterpart. The higher fraction of the hydrophobic molecular surface and the lower accessible area in cold adapted enzymes may also prevent the effects of cold denaturation that involves more favourable interactions between water and residues. Interestingly, special importance seems to be attached to conservation of structure around the beta-carbon of the amino acid (compare the frequencies of substitutions Thr $\rightarrow$ Val with Ser $\rightarrow$ Val, and Thr $\rightarrow$ Ile with Thr $\rightarrow$ Leu). Perhaps this helps to increase the flexibility in the geometry of the protein backbone.

In the core we find some different tendencies. There is significant increase in the number of charged and uncharged polar residues, and a decrease in the number of hydrophobic residues, although these trends are not very significant when we use the corresponding amino acid scales (where all residues may contribute) to measure the same. In total we may conclude that the psychrophilic population has an equal or less hydrophobic core. The residue weight increases, and consequently also volume, while the bulkiness decreases in the core region. Volume is important because of the ability of bigger residues to exclude water from the protein interior, to close internal cavities, and to decrease the entropic freedom of the unfolded protein backbone.

Flexibility

Catalysis often requires movements or ‘‘breathing’’ of all or of a particular region of the protein architecture. The ease of such molecular movement is considered important for cold adaptation. We observe a trend for more backbone flexibility at the molecular surface. In our analysis of secondary structures, it appears that psychrophilic proteins are likely to have more unfavourable secondary structure elements (Table 4), and which may destabilise the tertiary structure. The loop regions are also found to contain a decrease of inflexible Pro residues, as expected in a flexible structure, but show no significant increase of flexible Gly residues.

Flexibility is complex to infer. We used the new scale of peptide flexibility published by Huang and Nau (2003, 2005). *B*-values obtained from X-ray diffraction crystal-

lographic data provide measures of the average displacements of atoms in amino acid residues from their equilibrium geometry. These are derived from atomic vibrations on the millionth of a nanosecond (10^{-15} s) time scale and mainly reflect the shallowness of the potential centred round an energy minimum. The new scale, however, is based on a fluorazophore method for measuring the kinetics of end-to-end collision in polypeptides. Such collisions occur on the nanosecond to several hundred nanoseconds time scale, and should provide a more relevant biochemical measure of the conformational flexibility of the backbone. Both flexibility scales may be used to represent motion of the backbone, but at different time resolutions. Molecular dynamics simulations are typically done over a few nanoseconds.

In Tables 6 and 7, it is also of importance to note that the shape property, defined as position of branch point in the side chain (e.g. ranging from 0 in Alanine to 5 in Argenine), decreases at the molecular surface with adaptation to cold. The opposite trend is observed for Gibbs free energy change of hydration for native protein. Both these observations are direct extensions in agreement with results found earlier in a dataset of 14 thermophilic protein families (Gromiha et al. 1999). A decrease in shape may indicate a more flexible exterior because of early or no branching, and an increase in Gibbs free energy change will theoretically decrease the stability by decreasing the exposure of polar atoms for native proteins (Gromiha et al. 1999).

Some of the observations reported above, may possibly be an artefact of the choice of basing the prediction of secondary and spatial structures on a mesophlic sequence (*V. cholerae*), and not a psychrophilic sequence. However, the predictors we used (Adamczak et al. 2004, 2005) are train on structures from the PDB database, which is almost totally mesophilic.

The handling of gaps in the alignment analysis is problematic, as gaps do not exist in real structures, and it is also very difficult to get the correct position of gaps in multiple alignments. The missing physicochemical quantification of gaps represents an additional problem, because the statistical analysis has to ignore data in positions with gaps. Obviously, gap regions may contain valuable information, so a different strategy should be looked for.

Paralogs (if any) are included in the present analysis, because temperature adaptation may involve both copies. But paralogs can also induce a problem, since an additional adaptation (to new functionality) may potentially mask out the cold adaptation. Identifying such paralogs will be more correct and appropriate when the whole genome of *V. salmonicida* is completed.

Even though this study is based on new data from closely related proteins, and the number of occurrences of over 20% of the $x \rightarrow y$ replacements exceeds 100, limitations remain. Still with this amount of data, some of the amino acid substitutions are sampled only a few times, and additional data will be required to confirm or to refute some of the observed directional trends. Although this is unlikely to affect the overall trends that we discussed, it does limit discovery of more particular changes that might prove to be critical in specific contexts. Increasing the data set will better address these issues and will allow us to find out whether some trends are dependent on specific organisms, proteins, structural or cellular contexts.

One of the factors, which may affect amino acid composition, is the variation in genome GC content, ranging from 39–41% in *V. fischeri* and 47–49% in *V. cholerae* (Garrity 2005). This may not be regarded as very dissimilar, and in addition adaptations are considered to take place both at the DNA, RNA and protein level (Hickey and Singer 2004). But it is difficult to assess how much is a primary effect and how much is secondary or tertiary, and recent studies have concluded that in general there are no significant evidence of relationship between optimal temperature and GC content (Lambros et al. 2003; Wang et al. 2006).

Conclusions

Our comparative genome analyses suggest that adaptation to cold seems to be obtained through many minor structural modifications rather than certain amino acids substitution, and is usually obtained by exchanging a number of amino acids in comparison with its mesophile ortholog. To demonstrate that the observed differences are involved in the thermostability, the mesophilic protein dataset must also be compared with other mesophilic orthologs to remove noise or background differences. This kind of strategy may be used as a reasonable statistical model in all types of studies that utilise the results of comparison between two datasets to extract effective factors of temperature adaptation or other features of proteins.

We have applied and expanded the methods of comparative analysis of protein compositions, substitutions and property patterns; and we were able to detect several significant strand-specific factors affecting the cold adaptation in a dataset of 298 *Vibrionaceae* protein families. Several chemical properties were significantly correlated with cold adaptation; and most of the strongest correlations are found at the molecular surface.

Our results confirm the customary hypothesis of preference for more flexible amino acids at the molecular surface. Our analyses also reveal that psychrophilic and mesophilic proteins have similar hydrophobic and charge contribution to the interior, while the surface region of

psychrophile proteins is significantly more hydrophobic. For intracellular proteins a significant difference in the distribution of surface charge residues are also observed, and results in a more negative electrostatic potential, which is considered beneficial for the solubility of the protein in cold water. In terms of amino acids, Ile, Asn, Ala and Gln are found to have the most psycrophilic involvement. Ile and Asn are enhanced whereas Gln and Ala are suppressed. Their overall effect is primarily to increase external hydrophobicity as well as destabilising the secondary structure elements. Some factors normally associated with thermal adaptation, such as lower Pro content in loops, are compatible with our study, while higher Gly content was not found.

As a general comment, our results must be regarded with care. Although several factors have been shown to contribute to thermal adaptation, not all factors contribute to thermal adaptation in a given protein. The general strategy mentioned above may just be considered suitable for an ''average'' psychrophilic protein.

A second concern is that although the current sequence data are restricting to the family of *Vibrionaceae* and permitted the analysis of large numbers of very similar sequences, there still exist other environmental factors or characteristics of the organisms that might have affected the observed amino acid substitutions. One obvious factor is that *P. profundum* also is a moderate piezophile (pressure-loving) organism adapted to deep-sea life, although experimental data suggest that its proteins are not exclusively adapted to high pressure (Vezzi et al. 2005). It would be advantageous to verify the observed trends by studying other groups of closely related mesophilic and psychrophilic organisms. To improve the sample size, we are generating additional sequence data from cold adapted organisms.

Acknowledgments The present study was supported by the National Program for Research in Functional Genomics in Norway (FUGE). The Matlab toolbox DeltaProt (http://www.math.uit.no/bi/deltaprot/) was used for the data analyses.

References

Adamczak R, Porollo A, Meller J (2004) Accurate prediction of solvent accessibility using neural networks based regression. Proteins Struct Funct Bioinf 56:753–767

Adamczak R, Porollo A, Meller J (2005) Combining prediction of secondary structure and solvent accessibility in proteins. Proteins Struct Funct Bioinf 59:467–475

Argos P, Rossmann MG, Grau UM et al (1979) Thermal stability and protein structure. Biochemistry 18(25):5698–5703

Bae E, Phillips GN (2004) Structures and analysis of highly homologous psychrophilic, mesophilic, and thermophilic adenylate kinases. J Biol Chem 279(27):28202–28208

Britton KL, Baker PJ, Borges KMM et al (1995) Insights into thermal stability from a comparison of the glutamate dehydrogenases from *Pyrococcus furiosus* and *Thermococcus litoralis*. Eur J Biochem 229:688–695

Campanaro S, Vezzi A, Vitulo N et al (2005) Laterally transferred elements and high pressure adaptation in Photobacterium profundum strains. BMC Genomics 6:122

Chakravarty S, Varadarajan R (2000) Elucidation of determinants of protein stability through genome sequence analysis. FEBS Lett 470:65–69

Chakravarty S, Varadarajan R (2002) Elucidation of factors responsible for enhanced thermal stability of proteins: a structural genomics based study. Biochemistry 41(25):8152–8161

Chou PY, Fasman GD (1978) Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol 47:45–148

Colquhoun DJ, Alvheim K, Dommarsnes K et al (2002) Relevance of incubation temperature for Vibrio salmonicida vaccine production. J Appl Microbiol 92(6):1087–1096

Creighton TE (1991) Stability of folded conformations. Curr Opin Struct Biol 1:5–16

D'Amico S, Claverie P, Collins T et al (2002) Molecular basis of cold adaptation. Philos Trans R Soc Lond Ser B Biol Sci 357(1423):917–924

Daniel RM, Dines M, Petach HH (1996) The denaturation and degradation of stable enzymes at high temperatures. Biochem J 317:1–11

Fasman GD (ed) (1976) Proteins. Handbook of biochemistry and molecular biology. CRC Press, Cleveland

Fields PA (2001) Review: protein function at thermal extremes: balancing stability and flexibility. Comp Biochem Physiol Part A Mol 129(2–3):417–431

Fleiss JL et al (2003) Statistical methods for rates and proportions, 3rd edn. Wiley series in probability and statistics, New York

Fukui T, Atomi H, Kanai T et al (2005) Complete genome sequence of the hyperthermophilic archaeon Thermococcus kodakaraensis KOD1 and comparison with Pyrococcus genomes. Genome Res 15(3):352–363

Garrity GM (2005) Bergey's manual of systematic bacteriology, vol 2B, 2nd edn. Plenum, US

Georlette D, Blaise V, Collins T et al (2004) Some like it cold: biocatalysis at low temperatures. Fems Microbiol Rev 28(1):25–42

Gianese G, Argos P, Pascarella S (2001) Structural adaptation of enzymes to low temperatures. Protein Eng 14(3):141–148

Gianese G, Bossa F, Pascarella S (2002) Comparative structural analysis of psychrophilic and meso- and thermophilic enzymes. Proteins 47(2):236–249

Gromiha MM, Oobatake M, Sarai A (1999) Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophys Chem 82(1):51–67

Gunsteren WF, Mark AE (1992) Prediction of the activity and stability effects of site directed mutagenesis on a protein core. J Mol Biol 227:389–395

Haney PJ, Badger JH, Buldak GL et al (1999) Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic Methanococcus species. PNAS 96(7):3578–3583

Heidelberg JF, Eisen JA, Nelson WC et al (2000) DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae. Nature 406:477–483

Hickey DA, Singer GAC (2004) Genomic and proteomic adaptations to growth at high temperature. Genome Biol 5(10):article 117

Huang F, Nau WM (2003) A conformational flexibility scale for amino acids in peptides. Angew Chem Int Ed 42:2269–2272

Huang F, Nau WM (2005) Photochemical techniques for studying the flexibility of polypeptides. Res Chem Intermed 31(7–8):717–726

Karlin S, Brocchieri L, Trent J, Blaisdell BE, Mrazek J (2002) Heterogeneity of genome and proteome content in bacteria, archaea, and eukaryotes. Theor Popul Biol 61:367–390

Kazuoka T, Masuda Y, Oikawa T et al (2003) Thermostable aspartase from a marine psychrophile, Cytophaga sp KUC-1: molecular characterization and primary structure. J Biochem 133(1):51–58

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157(1):105–132

Lambros RJ, Mortimer JR, Forsdyke DR (2003) Optimum growth temperature and the base composition of open reading frames in prokaryotes. Extremophiles 7(6):443–450

Levitt M (1976) Simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 104(1):59–107

Lim WA, Sauer RT (1991) The role of internal packing interactions in determining the structure and stability of a protein. J Mol Biol 219(2):359–376

Lovell SC, Word JM, Richardson JS, Richardson DC (2000) The penultimate rotamer library. Proteins 40(3):389–408

Makino K, Oshima K, Kurokawa K et al (2003) Genome sequence of Vibrio parahaemolyticus: a pathogenic mechanism distinct from that of *V. cholerae*. Lancet 361:743–749

Martin A, Sieber V, Schmid FX (2001) In-vitro selection of highly stabilized protein variants with optimized surface. J Mol Biol 309(3):717–726

Marx JC, Blaise V, Collins T et al (2004) A perspective on cold enzymes: current knowledge and frequently asked questions. Cell Mol Biol 50(5):643–655

McDonald JH, Grasso AM, Rejto LK (1999) Patterns of temperature adaptation in proteins from Methanococcus and Bacillus. Mol Biol Evol 16(12):1785–1790

Medigue C, Krin E, Pascal G et al (2005) Coping with cold: the genome of the versatile marine Antarctica bacterium Pseudoalteromonas haloplanktis TAC125. Genome Res 15(10):1325–1335

Methe BA, Nelson KE, Deming JW et al (2005) The psychrophilic lifestyle as revealed by the genome sequence of Colwellia psychrerythraea 34H through genomic and proteomic analyses. Proc Natl Acad Sci USA 102(31):10913–10918

Nishio Y, Nakamura Y, Kawarabayasi Y et al (2003) Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of Corynebacterium efficiens. Genome Res 13(7):1572–1579

Notredame C, Higgins DG, Heringa J (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 302:205–217

Oobatake M, Ooi T (1993) Hydration and heat-stability effects on protein unfolding. Prog Biophys Mol Biol 59(3):237–284

Pe'er I, Felder CE, Man O, Silman I, Sussman JL, Beckmann JS (2004) Proteomic signatures: amino acid and oligopeptide compositions differentiate among phyla. Proteins 54(1):20–40

Pollastri G, Baldi P, Fariselli P et al (2002) Prediction of coordination number and relative solvent accessibility in proteins. Proteins 47(2):142–153

Ponnuswamy PK, Gromiha MM (1993) Prediction of transmembrane helices from hydrophobic characteristics of proteins. Int J Pept Protein Res 42:326–341

Rabus R, Ruepp A, Frickey T et al (2004) The genome of *Desulfotalea psychrophila*, a sulfate-reducing bacterium from permanently cold Arctic sediments. Environ Microbiol 6(9):887–902

Richards FM (1977) Areas, volumes, packing and protein structure. Annu Rev Biophys Bioeng 6:151–176

Ruby EG, Urbanowski M, Campbell J et al (2005) Complete genome sequence of *Vibrio fischeri*: a symbiotic bacterium with pathogenic congeners. Proc Natl Acad Sci USA 102:3004–3009

Sadeghi M, Naderi-Manesh H, Zarrabi M et al (2006) Effective factors in thermostability of thermophilic proteins. Biophys Chem 3:256–270

Saunders NFW, Thomas T, Curmi PMG et al (2003) Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea *Methanogenium frigidum* and *Methanococcoides burtonii*. Genome Res 13(7):1580–1588

Taylor WR (1986) The classification of amino acid conservation. J Theor Biol 119:205–218

Thompson FL, Iida T, Swings J (2004) Biodiversity of vibrios. Microbiol Mol Biol Rev 68(3):403–431

Thorvaldsen S, Ytterstad E, Flå T (2006) Property-dependent analysis of aligned proteins from two or more populations. In: Jiang T, Yang UC, Chen YPP, Wong L (eds) Proceedings of the 4th Asia-Pacific bioinformatics conference. Imperial College Press, pp 169–178

van Passel MWJ, Bart A, Thygesen HH et al (2005) An acquisition account of genomic islands based on genome signature comparisons. BMC Genomics 6:Art. No. 163

Vezzi A, Campanaro S, D'Angelo M et al (2005) Life at depth: Photobacterium profundum genome sequence and expression analysis. Science 307:1459–1461

Wang HC, Susko E, Roger AJ (2006) On the correlation between genomic G + C content and optimal growth temperature in prokaryotes: data quality and confounding factors. Biochem Biophys Res Commun 342(3):681–684

Wieslander Å, Rosén M (2002) Cell membranes and transport. In: Razin S, Herrmann R (eds) Molecular biology and pathogenicity of mycoplasmas. Kluwer, Dordrecht

Yu CS, Lin CJ, Hwang JK (2004) Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci 13:1402–1406

Zimmerman JM, Eliezer N, Simha R (1968) The characterization of amino acid sequences in proteins by statistical methods. J Theor Biol 21(2):170–201