Available online at www.sciencedirect.com
ScienceDirect
Protein Expression and Purification 57 (2008) 234–243
Protein Expression & Purification
www.elsevier.com/locate/yprep

Addition of a polypeptide stretch at the N-terminus improves the expression, stability and solubility of recombinant protein tyrosine phosphatases from *Drosophila melanogaster*

Lalima L. Madan, B. Gopal *

Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

Received 10 August 2007, and in revised form 1 October 2007
Available online 13 October 2007

Abstract

The production of recombinant proteins in *Escherichia coli* involves substantial optimization in the size of the protein and over-expression strategies to avoid inclusion-body formation. Here we report our observations on this so-called construct dependence using the catalytic domains of five *Drosophila melanogaster* receptor protein tyrosine phosphatases as a model system. Five strains of *E. coli* as well as three variations in purification tags viz., poly-histidine peptide attachments at the N- and C-termini and a construct with Glutathione-*S*-transferase at the N-terminus were examined. In this study we observe that inclusion of a 45 residue stretch at the N-terminus was crucial for over-expression of the enzymes, influencing both the solubility and the stability of these recombinant proteins. While the addition of negatively charged residues in the N-terminal extension could partially rationalize the improvement in the solubility of these constructs, conventional parameters like the proportion of order promoting residues or aliphatic index did not correlate with the improved biochemical characteristics. These findings thus suggest the inclusion of additional parameters apart from rigid domain predictions to obtain domain constructs that are most likely to yield soluble protein upon expression in *E. coli*.

Keywords: Protein tyrosine phosphatase; Expression; Inclusion bodies; Protein solubility; Light scattering

Among the various host expression systems available for the large-scale production of recombinant proteins, *Escherichia coli* is an attractive choice. The primary advantage of *E. coli* as a host expression system is the availability of specifically engineered strains and amino acid auxotrophs that can be used for recombinant protein production. The major drawback, however, is the inability to effectively couple post-translational processes with recombinant protein expression and a commonly encountered scenario wherein a substantial proportion of the recombinant protein is lost in insoluble aggregates referred to as inclusion bodies. While there does not appear to be a convenient method to compensate for the post-translation machinery that is absent in *E. coli*, several sequence and conformational determinants have been proposed to predict inclusion body formation of recombinant proteins. These include specific amino acid occurrences at the N- or C-termini [1–5], presence of fusion tags [5–7] or features in the primary sequence and the size of the recombinant protein [8,9]. Optimization of recombinant protein production based on these empirical features thus involves substantial experimentation with the size of the recombinant protein as well as expression conditions. Here we report our observations on the influence of the construct on the expression and *in vitro* biochemical characteristics of recombinant proteins based on the case study of five protein tyrosine phosphatases (PTPs)[1] from *Drosophila melanogaster*.

* Corresponding author. Fax: +91 80 23600535.
E-mail address: bgopal@mbu.iisc.ernet.in (B. Gopal).

[1] *Abbreviations used:* PTP, protein tyrosine phosphatase; pNPP, para-nitrophenyl phosphate; IPTG, isopropylthioglucopyranoside; GST, glutathione *S*-transferase.

1046-5928/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.pep.2007.10.002

Five *D. melanogaster* PTPs (DLAR, DPTP10D, DPTP69D, DPTP52F and DPTP99A) are selectively expressed on the central nervous system (CNS) axons and growth cones in the *Drosophila* embryo. These five proteins share similar sequence features—an extracellular domain composed of Immunoglobulin and Fibronectin-III repeats, a short membrane-spanning segment and a cytosolic region that has either one (DPTP10D and DPTP52F) or two (DLAR, DPTP99A and DPTP69D) PTP domains (Fig. 1). The catalytic domains of these proteins dephosphorylate tyrosine residues of their cognate substrate proteins and regulate multiple signal transduction processes during *Drosophila* development [10]. In an effort to obtain suitable quantities of protein to characterize the *Drosophila* PTPs *in vitro*, we proposed to obtain recombinant proteins using a prokaryotic expression system. The recombinant proteins were expressed as fusion constructs with either N- or C-terminal poly-histidine tags or with the glutathione *S*-transferase protein (GST) attached to the N-terminus. The presence of the PTPs in the soluble form was examined by an activity assay using para-nitrophenyl phosphate (pNPP) as a substrate and by Western blot analysis using antibodies raised against the affinity tags fused to the recombinant proteins. Initial efforts to purify the catalytic domains revealed that the recombinant *Drosophila* PTPs formed inclusion bodies when over-expressed in *E. coli*. Optimization trials to maximize the amount of recombinant protein in the soluble fraction suggested that an N-terminal polypeptide extension of ca 45 residues was crucial for this process. The improvement in the yield of the recombinant PTPs was independent of the affinity tags attached to the catalytic domains. Although this study is confined to members of the PTP family, the results reported in this manuscript highlights the need to incorporate additional sequence and conformational features apart from conventional domain definitions in the design of gene constructs for the expression of recombinant proteins in *E. coli*.

Materials and methods

Sequence analysis of the five PTP domains from D. melanogaster

The sequences of all the five PTPs viz., DLAR, DPTP69D, DPTP99A, DPTP10D and DPTP52F were retrieved from Flybase (FlyBase@flybase.bio.indiana.edu). The Flybase annotations for DPTP52F, DPTP10D, DLAR, DPTP99A and DPTP69D are CG18243, CG1817, CG10443, CG2005 and CG10975, respectively. The domain definitions from the protein sequences were obtained using the NCBI conserved domain search

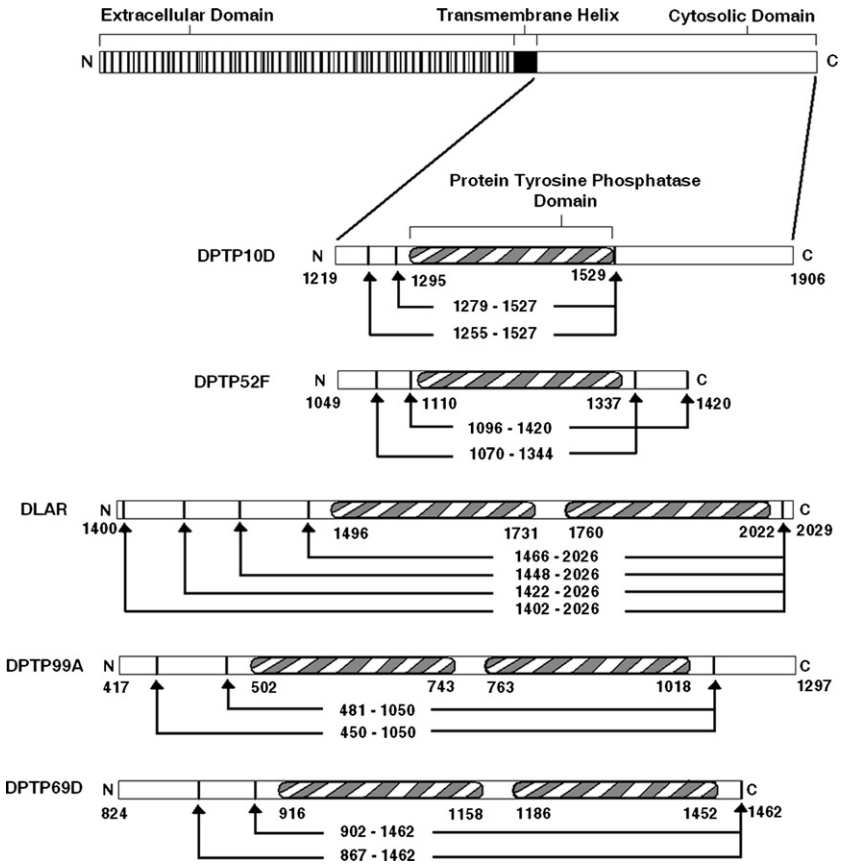

Fig. 1. Domain arrangement in the five protein tyrosine phosphatases of *D. melanogaster*. The boundaries of the different expression constructs used in this study can be identified by arrow marks.

(http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) and 'The Eukaryotic Linear Motif resource for functional sites in proteins' web-server (http://elm.eu.org/links.html). The membrane spanning segments of the proteins were identified by submitting the sequences to the TMHMM Server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM-2.0/) and PRED-TMR (http://o2.db.uoa.gr/PRED-TMR/). Multiple sequence alignments were performed on the Multalin Server (http://prodes.toulouse.inra.fr/multalin/multalin.html).

Expression constructs

Cosmids containing the full-length receptor protein tyrosine phosphatases (RPTPs) were obtained from Prof. K. Zinn (Caltech). Oligonucleotide primers used in the PCR amplification were obtained from Sigma–Aldrich Co., USA. The expression vectors pET15b and pET22b (Novagen) and pGEX4T1 (Amersham-Pharmacia, Inc.,) were used in the cloning experiments. These vectors were modified to accommodate NheI and XhoI restriction sites and all the constructs reported in this study were inserted between the NheI and XhoI sites from the 5′- and 3′-position. The PTP catalytic domains were PCR amplified from the full-length constructs present in the cosmids. The positive clones were screened by colony PCR and insert release assays. The sequences of all the clones were confirmed by single primer extension sequencing (Macrogen Inc., South Korea).

Expression and analysis of the recombinant proteins

The plasmids were transformed into different strains of *E. coli* by the heat shock method and transformants were then grown in LB broth with 100 μg/ml ampicillin as an antibiotic for selection. The *E. coli* strains BL21(DE3), B834(DE3), BL21 C41(DE3), Rosetta (DE3) and Rosetta-gami (DE3) strains were used in this study. For recombinant protein expression, the cells were grown to an OD of 0.6 at 600 nm at 37 °C and induced with 1 mM IPTG. The culture was then allowed to grow to an OD of 2.0 and harvested at 12,000 rpm for 10 min. The cells were lysed by sonication in lysis buffer (50 mM HEPES, pH 8.0, 100 mM NaCl, 2.5 mM EDTA, 5 mM DTT and 1 mM PMSF) and centrifuged to remove the cell debris. Both the pellet and the supernatant were loaded onto an SDS–PAGE to check for protein expression and solubility. Different variations in temperature were tried after induction viz., 37, 25, 17 and 12 °C. The IPTG concentration was also optimized by variations from 0.1 mM to 1 mM. The concentration of IPTG finally used for induction was 0.1 mM and the temperature was lowered to 12 °C post-induction. The extent of the recombinant protein in the soluble fraction was estimated by an assay for phosphatase activity in the crude extract in addition to its profile on a SDS–PAGE. For the GST-fusion constructs, both the supernatant and pellet obtained after sonication were probed using HRP conjugated rabbit anti-GST IgG antibody (Bangalore Genei, Co.). A dilution of 1:2000 of the antibody was used for the Western blot experiments.

Phosphatase activity using para-nitro phenyl phosphate

The supernatant obtained after sonication and centrifugation was tested for phosphatase activity using para-nitro phenyl phosphate (pNPP) as the substrate. The reaction buffer used was 50 mM HEPES pH 7.2 containing 100 mM NaCl, 2.5 mM EDTA and 5 mM DTT. The reaction mixture contained 100 μg/μl bovine serum albumin (BSA), 5 mM pNPP and 20 μl of the lysate. The reaction mixture was incubated at 37 °C for 15 min and the reaction was stopped by the addition of 2 N NaOH. The product formed was calculated by taking fixed wavelength measurements at 405 nm. Enzyme activity was calculated using the molar extinction coefficient of pNPP at alkaline pH as 17,800 M^{-1} cm^{-1}. Both the induced and the un-induced whole cell lysates were analyzed. The results of the phosphatase activity were compared with two control experiments using the substrate alone as well as lysates of the non-transformed competent cells grown under identical conditions to negate the contribution of *E. coli* phosphatases for the reaction. The protein concentration of the cell free lysate was obtained using a standard Bradford protein estimation assay.

Purification of DPTP10D (1279–1527) and DPTP10D (1255–1527)

As the construct lacking the N-terminal peptide went into inclusion bodies, it was purified by denaturation followed by refolding using a modification of the protocol reported earlier [11,12]. A 3 L culture of BL21 (DE3) cells induced for the over-expression of DPTP10D (1279–1527) in the pET22b vector was harvested and 14 g of dry cell pellet was resuspended in 50 ml of phosphate buffered saline. The cells were lysed by sonication and centrifuged at 12,000 rpm to separate the pellet and supernatant. The supernatant was discarded and the pellet washed three times with 50 ml of 100 mM Tris, pH 8.0, 250 mM NaCl, 5 mM EDTA, 5 mM DTT, 2 mM PMSF and 0.5% Triton X-100. The final wash was with the same buffer lacking the detergent. The pellet was then dissolved in a minimal amount of denaturation buffer (5 ml) containing 6 M guanindine HCl, 50 mM Tris, pH 8.0, 150 mM NaCl, 10 mM EDTA, 10 mM DTT and 2 mM phenyl methyl sulphonyl fluoride (PMSF). The pellet was allowed to dissolve overnight at 4 °C with constant mixing by an end-to-end rotator. The pellet was then centrifuged at 25,000*g* for 45 min to remove the un-dissolved debris. The supernatant obtained was refolded by rapid dilution in refolding buffer containing 100 mM HEPES, pH 8.0, 150 mM NaCl, 1 M arginine, 25% sucrose, 10 mM EDTA and 2 mM PMSF. One millimolar oxidized glutathione and 10 mM reduced glutathione were used as the redox couple. The solution

containing the denaturated protein was then diluted 1000 times such that the final guanidine concentration reached 0.006 M. The protein was left to stir in the refolding buffer at 4 °C for 18 h. The refolded protein was concentrated under nitrogen pressure using an Amicon Stir cell with a 10 kDa exclusion membrane (AMICON, USA). The concentration of the protein was estimated by the absorbance at 280 nm ($\varepsilon_{280} = 48{,}400\ M^{-1}\ cm^{-1}$).

The longer construct harboring the extra amino acids at the N-terminal was purified in its soluble form. Three liters of BL21(DE3) cells expressing the DPTP10D (1255–1527) in pET22b vector were harvested by centrifugation at 6000 rpm at 4 °C. Twelve grams of dry cell pellet was dissolved in 35 ml of PBS buffer sonicated to lyse the cells. The lysate was centrifuged at 12000 rpm to remove the cell debris. The supernatant was loaded onto 5 ml of Ni-NTA beads (His-Select, Sigma–Aldrich) to aid purification with the help of the poly-histidine tag at the C-terminus. The protein was eluted in PBS containing 100 mM imidazole. The eluate was analyzed by SDS–PAGE and passed through a Sephacryl S-200 size exclusion column (Amersham-Pharmacia, Inc.) equilibrated with 50 mM Tris, pH 8.0, 250 mM NaCl, 10 mM EDTA and 2 mM PMSF. An aliquot of the protein to be used for temperature dependent aggregation studies was dialyzed against the refolding buffer as mentioned for the smaller construct. The protein concentration was ascertained by measuring the absorbance at 280 nm ($\varepsilon_{280} = 49890\ M^{-1}\ cm^{-1}$).

Temperature dependent aggregation of DPTP10D (1279–1527) and DPTP10D (1255–1527)

Both the DPTP10D constructs weree diluted in 50 mM HEPES, pH 8.0, 150 mM NaCl such that 2 μM of each variant was examined for temperature dependent aggregation. Both the protein solutions were scanned for scattering between 340 and 700 nm using a UV–Visible Spectrophotometer (JASCO, Japan). The buffer alone was scanned between the same wavelengths and temperature range in a control experiment. These scans were repeated at 403 nm between 5 and 90 °C by increasing the temperature at a rate of 2 °C/min to obtain a thermal aggregation curve. Turbidity of the solution was calculated as $\tau = 2.303\ (\text{Absorbance}_{403nm})/l$ [13], where l is the path length of the cuvette used. The minimum and maximum turbidity values were used as limits of 0 and 100% aggregation. The midpoint of inflexion was used to calculate the temperature of aggregation (T_{agg}) of the protein. The data were analyzed using Sigma Plot software (Omega Scientific, Inc.).

Analysis of the N-terminal peptide

The amino acid composition of the N-terminal region and the proteins were obtained using the ExPASy PROTPARAM web interface http://www.expasy.ch/tools/protparam.html. The secondary structure propensities were obtained using PSIPRED [14]. The definitions of order-promoting amino acids (Asn, Cys, Ile, Leu, Phe, Trp, Tyr and Val), disorder-promoting amino acids (Ala, Arg, Gln, Glu, Lys, Pro, Gly and Ser) and neutral amino acids (Asp, His, Met and Thr) were based on earlier observations [15]. The folding probability of the construct was estimated by submitting the sequence to the FoldIndex server (http://bip.weizmann.ac.il/fldbin/findex). The aliphatic index of the constructs used was calculated as reported earlier [8]. The structures of the representative members were downloaded from the protein data Bank (http://www.rcsb.org) and the structures were viewed using the Pymol software (DeLano-scientific LLC).

Results and discussion

Analysis of different RPTP constructs for their expression in E. coli

The strategy to clone and express the recombinant PTPs involved inserting the gene of interest into the pET15b expression vector (yielding a recombinant protein with a poly-histidine tag at the N-terminus, Novagen, Inc.), pET22b (recombinant protein with a poly-histidine tag at the C-terminus, Novagen, Inc.) and pGEX4T1 (fusion protein with glutathione *S*-transferase at the N-terminus, Amersham-Pharmacia, Inc.). The use of different fusion tags and *E. coli* expression strains was necessary as the choice of a fusion system may affect the solubility of the construct [6]. The initial set of expression constructs were designed to satisfy the catalytic PTP domain definition as obtained from the NCBI domain search and inputs from the eukaryotic linear motif search for functional sites in proteins (Fig. 1). All the constructs contained the 10 conserved sequence-structure motifs (M1–M10) of a PTP domain [16]. In these constructs, the motif M1 formed the N-terminal domain boundary ending with the motif M10 at the C-terminus. Over-expression of these five catalytic domain constructs viz., DLAR (1466–2026), DPTP69D (902–1462), DPTP99A (481–1050), DPTP10D (1279–1527) and DPTP52F (1096–1344) showed that all of them formed insoluble inclusion bodies. Inclusion body formation occurred regardless of the choice of the expression vector or the *E. coli* strains that were examined. Thus while the enzyme assay showed no phosphatase activity in the crude cell lysate (Fig. 2), Western blots using the glutathione *S*-transferase (GST) antibody showed that the recombinant proteins formed inclusion bodies and were not present in the soluble fraction (Figs. 2 and 3). Conventional strategies to improve upon the solubility of these recombinant proteins with variations in the temperature of growth after induction with isopropylthioglucopyranoside (IPTG) or changing IPTG concentrations did not result in any improvement in the solubility of the recombinant proteins. On the basis of reports that suggested that a polypeptide stretch preceding motif M1 is involved in the regulation of enzyme activity in the case of a few members of the PTP family [17,18], a new set of expression con-

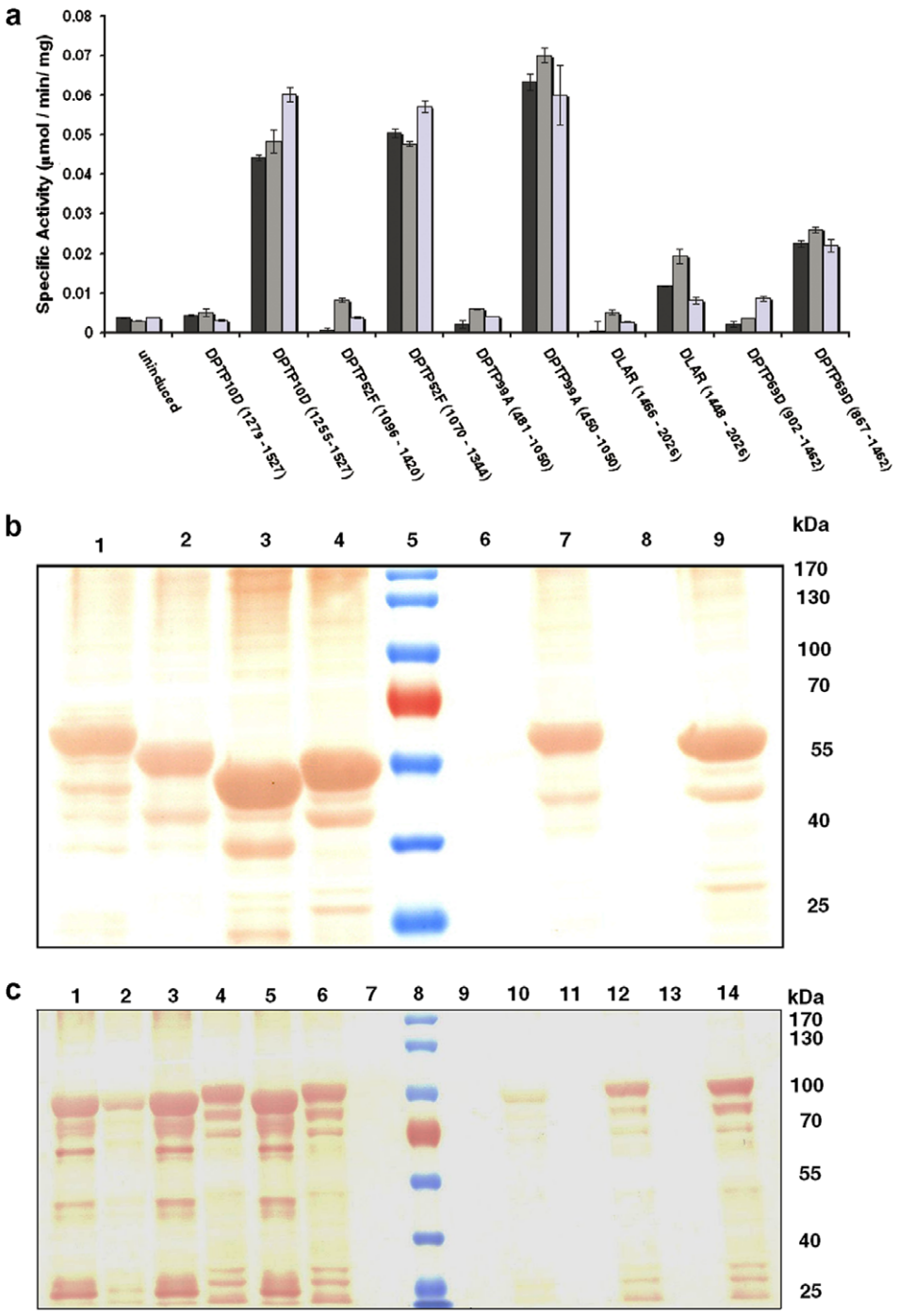

Fig. 2. (a) Phosphatase assay for crude *E. coli* extracts containing the recombinant PTPs. Lysates of cells expressing the constructs with the N-terminal extension show a marked increase in phosphatase activity when compared to the classical PTP constructs or the cell extracts alone. This bar chart shows the pET15b construct (black), pET22b construct (grey) and the pGEX4T1 construct (blue). (b) Western blot analysis to probe for the presence of single domain PTPs fused to Glutathione *S*-transferase in the soluble fractions or precipitate (inclusion bodies). The presence of soluble protein is characterized by a sharp increase in phosphatase activity (panel a) as well as a band corresponding to the fusion protein in the supernatant of the *E. coli* lysate (Lanes 7 and 9). Lane 1: pellet DPTP52F (1096–1420, short construct), Lane 2: pellet DPTP52F (1070–1344, extended construct), Lane 3: pellet DPTP10D (1279–1527, short construct), Lane 4: pellet DPTP10D (1255–1527, extended construct), Lane 5: molecular weight marker, Lane 6: Supernatant DPTP52F (1096–1420), Lane 7: Supernatant DPTP52F (1070–1344), Lane 8: Supernatant DPTP10D (1279–1527), Lane 9: Supernatant DPTP10D (1255–1527). (c) Western blot analysis to probe for the GST fusion double domain PTPs in the insoluble and supernatant fractions. Lane 1 : pellet DLAR (1466–2026, short construct), Lane 2: pellet DLAR (1448–2026,extended construct), Lane 3: pellet DPTP69D (902–1462, short construct), Lane 4: pellet DPTP69D (867–1462, extended construct), Lane 5: pellet DPTP99A (481–1050, short construct), Lane 6: pellet DPTP99A (450–1050, extended construct), Lane 7: Whole cell lysate BL21 un-transformed, Lane 8: molecular weight marker, Lane 9: Supernatant DLAR (1466–2026), Lane 10: Supernatant DLAR (1448–2026), Lane 11: Supernatant DPTP69D (902–1462), Lane 12: Supernatant DPTP69D (867–1462), Lane 13: Supernatant DPTP99A (481–1050), Lane 14: Supernatant DPTP99A (450–1050).

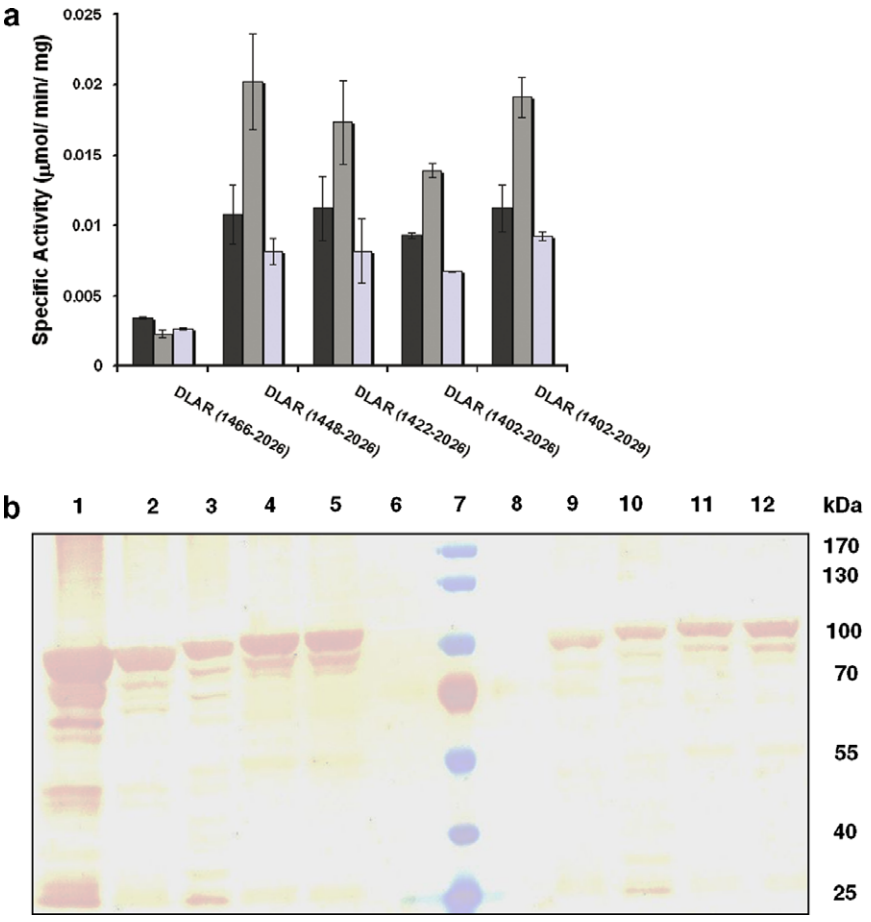

Fig. 3. (a) Phosphatase assay for crude *E. coli* extracts of cells expressing different lengths of DLAR. A construct with a 45 amino acid N-terminal extension is seen to be the most suited expression construct of DLAR. (b) Western blot analysis to probe for different constructs of GST fusion DLAR in the pellet and supernatant fractions. Lane 1: pellet DLAR (1466–2026), Lane 2: pellet DLAR (1448–2026), Lane 3: pellet DLAR (1422–2026), Lane 4: pellet DLAR (1402–2026), Lane 5: pellet DLAR (1402–2029), Lane 6: BL21 DE3 whole cell lysate, Lane 7: Molecular Weight Marker, Lane 8: Supernatant DLAR (1466–2026), Lane 9: Supernatant DLAR (1448–2026), Lane 10: Supernatant DLAR (1422–2026), Lane 11: Supernatant DLAR (1402–2026), Lane 12: Supernatant DLAR (1402–2029).

structs were designed to include a segment upstream of the classical PTP domain. These modified recombinant proteins thus had an N-terminal extension of 45 amino acids prior to the motif M1 of the PTP domain (Fig. 1). These five expression constructs viz., DLAR (1448–2026), DPTP99A (450–1050), DPTP69D (867–1462), DPTP10D (1255–1527) and DPTP52F (1070–1344) were cloned into all the three expression vectors described earlier. Interestingly, all these constructs yielded recombinant proteins that were soluble (Figs. 2 and 3). Furthermore, no substantial variation in the solubility was seen to suggest any preference for a specific fusion tag or an *E. coli* strain used for protein expression (Table 1).

The N-terminal polypeptide improves both, the stability and the solubility of the recombinant PTPs

The temperature dependence of protein aggregation and inclusion body formation has been extensively studied in a variety of model systems [19,20]. Indeed, an apparent correlation has been noted between these two factors whereby an increase in the temperature of aggregation (T_{agg}) suggests a lower propensity for a recombinant protein to form inclusion bodies [19–23]. In order to examine if this was the case with the *Drosophila* PTPs, two versions of DPTP10D were examined for their thermo-stability. This PTP was chosen because it is smaller in size when compared to other PTPs and could be refolded more easily than the other proteins. Thus DPTP10D (1279–1527) with a poly-histidine tag at the C-terminus was purified from inclusion bodies, whereas the extended PTP construct, DPTP10D (1255–1527), was purified from the soluble fraction. Protein aggregation with increase in temperature was measured by the increase in turbidity of the sample (Fig. 4, also see Supplementary Figs. 1 and 2). Both the variants of DPTP10D gave a biphasic plot when the aggregation was monitored as function of temperature at 403 nm (Fig. 4). The mid-point of inflection of the curves was used to calculate the temperature of aggregation T_{agg}. While two distinct ways of calculation of T_{agg} have been reported in literature [20,22], both show the propensity of a protein to form inclusion bodies to be linked to this parameter. DPTP10D (1279–1527) and DPTP10D (1255–1527) were found to have considerably different T_{agg} values, ca 46.4 °C in the

Table 1
Effect of different strains of *E. coli* on the expression of recombinant PTPs

RPTP	Vector	Expression in BL21 (DE3)	Expression in B834 (DE3)	Expression in BL21 C41 (DE3)	Expression in Rosetta (DE3)	Expression in Rosetta-gami (DE3)
DPTP10D (1279–1527)	pET15b	E P	E P	N	E P	E P
	pET22b	E P	E P	N	E P	E P
	pGEX4T1	E P	E P	E P	E P	E P
DPTP10D (1255–1527)	pET15b	E S	E S	E S	E S	E S
	pET22b	E S	E S	E S	E S	E S
	pGEX4T1	E S	E S	E S	E S	E S
DPTP52F (1096–1420)	pET15b	E P	E P	N	E P	E P
	pET22b	E P	E P	N	E P	E P
	pGEX4T1	E P	E P	EP	E P	E P
DPTP52F (1070–1344)	pET15b	E S	E S	E S	E S	E S
	pET22b	E S	E S	E S	E S	E S
	pGEX4T1	E S	ES	E S	E S	E S
DLAR (1466–2026)	pET15b	E S	E P	E P	E P	EP
	pET22b	E P	E P	E P	E P	E P
	pGEX4T1	E P	E P	E P	E P	E P
DLAR (1448–2026)	pET15b	E S	E S	E S	E S	E S
	pET22b	E S	E S	E S	E S	E S
	pGEX4T1	E S	E S	E S	E S	E S
DLAR (1422–2026)	pET15b	E S	E S	E S	E S	E S
	pET22b	E S	E S	E S	E S	E S
	pGEX4T1	E S	E S	E S	E S	E S
DLAR (1402–2026)	pET15b	E S	E S	E S	E S	E S
	pET22b	E S	E S	E S	E S	E S
	pGEX4T1	E S	E S	E S	E S	E S
DPTP99A (481–1050)	pET15b	E P	E P	E P	E P	E P
	pET22b	E P	E P	E P	E P	E P
	pGEX4T1	E P	E P	E P	E P	E P
DPTP99A (450–1050)	pET15b	E S	E S	E S	E S	E S
	pET22b	E S	E S	E S	E S	E S
	pGEX4T1	E S	E S	E S	E S	E S
DPTP69D (902–1462)	pET15b	E P	E P	E P	E P	E P
	pET22b	E P	E P	E P	E P	E P
	pGEX4T1	E P	E P	E P	E P	E P
DPTP69D (867–1462)	pET15b	E S	E S	E S	E S	E S
	pET22b	E S	E S	E S	E S	E S
	pGEX4T1	E S	E S	E S	E S	E S

The constructs lacking the N-terminal polypeptide extension formed inclusion bodies upon expression in all the strains examined here.
Expression of the recombinant PTP was determined by the presence of a band corresponding to the recombinant protein on SDS–PAGE gel followed by Western blot analysis as well as an assay for phosphatase activity using the whole cell lysate. The recombinant protein was considered soluble (E S) if it was present in the supernatant of the induced cell culture after sonication and showed phosphatase activity in the whole cell lysate. The insoluble protein (E P) is seen in the pellet after sonication of the induced culture and shows no phosphatase activity in the whole cell lysate. N indicates that the recombinant protein was not expressed in the *E coli.* strain.

case of DPTP10D (1279–1527) and 58.3 °C in the case of the extended construct, DPTP10D (1255–1527). The increase in the T_{agg} of DPTP10D (1255–1527) is thus in agreement with the lower propensity of this construct to form inclusion bodies.

Effect of known factors on the solubility of recombinant proteins in E. coli.—implications for the D. melanogaster PTPs

An inverse correlation between the net negative charge of a protein and its propensity to form inclusion bodies when over-expressed in *E. coli* has been noted earlier [24,25]. This feature is not evident in the case of the *Drosophila* PTPs (Supplementary Tables 1 and 2). For both DPTP10D and DPTP69D, the recombinant constructs with the N-terminal extension gave these proteins an additional negative charge of four units. For DPTP99A and DLAR the increments were by three and one unit(s), respectively. On the other hand, both the longer as well as the compact constructs of DPTP52F were estimated to have a net positive charge. However, if this analysis was restricted to the N-terminal polypeptide extension alone, we do observe a correlation with this empirical rule. As

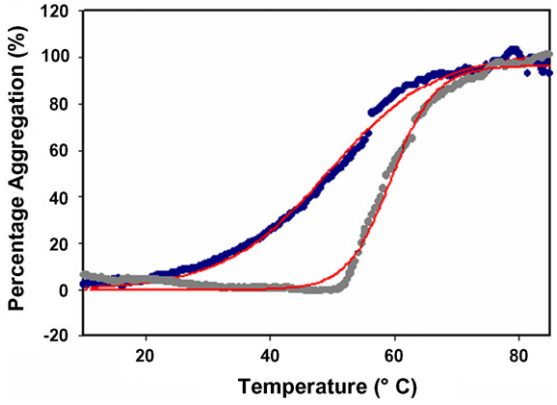

Fig. 4. Thermal denaturation studies on two constructs of DPTP10D. The two-state transition curve for the aggregation of DPTP10D (1279–1527) is shown in black and that for DPTP10D (1255–1527) is in grey. The sigmoidal curve fitted to the data is shown by the red line. The temperature of aggregation of DPTP10D (1279–1527) is 46.4 ± 0.4 °C while that of construct with the N-terminal extension, DPTP10D (1255–1527), is 58.3 ± 0.2 °C. For DPTP10D (1279–1527) the Absorbance at 403 nm at 0% aggregation was 0.0400 U and 0.1508 U at 100% aggregation. In the case of DPTP10D (1255–1527) the absorbance at 403 nm at 0% aggregation was 0.0007 U and 0.1566 U at 100% aggregation (see supplementary figure 1).

expected, we also observe an upper limit to this strategy. Of the three constructs of DLAR, for example, the longest DLAR (1422–2026) construct with an additional 75 residues at the N terminus is less soluble than the protein with a 45 residue N-terminal extension (Fig. 3).

Another sequence parameter linked with the stability of a protein is the aliphatic index (AI)—the mole fraction of Ala, Val, Ile and Leu residues found in a protein. As the AI of proteins from thermophilic organisms are significantly higher than mesophiles, the AI is often considered a measure of thermal stability of the protein [26]. The positive correlation of the mean AI of soluble and insoluble proteins over-expressed in *E. coli* have also been reported [8]. In the case of the PTPs, however, the N-terminal polypeptide extension does not substantially alter the AI (details compiled in supplementary Table 3). A comparison of the disordered residue content in the N-terminal extension of the PTPs provides an even more surprising feature. The theoretical number of the so-called disorder promoting residues actually increases in the DPTP10D, DPTP52F and DPTP99A longer constructs whereas it remains unchanged in the case of DLAR and DPTP69D. This correlates with the results obtained from folded-ness prediction algorithms such as Fold Index [27] that suggest that the N-terminal region of the longer constructs is likely to be disordered. Notwithstanding these contradictory observations for the PTPs, the positive effect of the N-terminal stretch on the solubility of the recombinant proteins upon over-expression is not without precedent. In the case of the chimeric human liver aldehyde dehydrogenases, for example, the cytosolic and mitochondrial isozymes share only 15% identity in the first 21 residues (whereas the overall identity is closer to 70%). This N-terminal region in the mitochondrial isozyme could solubilize the protein, whereas point mutations were needed in the corresponding region of the cytosolic isozyme to yield soluble protein when over-expressed in *E. coli* [28].

Secondary structure analysis of the N-terminal extension

The polypeptide that precedes the M1 motif of the PTP domain is involved in the dimerization of some PTP domains leading to its inactivation [17,29] while forming secondary binding sites for the substrate peptide in others [30]. In the case of RPTPα [17], secondary structural elements referred to as helix α1' and helix α2' were identified in the crystal structure. Given the low sequence similarity in this region between members of the PTP family, it was not surprising to observe that these α-helices vary substantially in length and also in their position from the M1 Motif of the catalytic domain (Fig. 5). In the cases where the N-terminal polypeptide contains α-helices, helix α1' varies from a maximum span of 17 residues in domain 1 of PTPγ (PDB code: 2H4V) to a minimum of 7 residues in HePTP (2A3K) and PTP1B (2HNP). The helix α2' varies from 19 residues in domain 1 of PTPγ to eight residues in PTPH1 (2B1J). The location of these helices from the catalytic domain shows considerable diversity as well from 50 residues spacing between helix α1' and the M1 motif in SHP2 (2SHP) to 35 residues in the case of PCPTP1 (2A8B), PTPSL/BR7 (1JLN) and HePTP (2A3K). Secondary structure prediction using PSIPRED [14] suggested the presence of two α helices in the N-terminal region of all the *Drosophila* PTPs examined in this study. An estimate of the optimum length of the polypeptide extension was obtained using expression constructs of DLAR. Three constructs of the DLAR were examined, one with 45 additional amino acids DLAR (1448–2026) prior to motif M1, a longer version with 75 amino acids at the N-terminal DLAR (1422–2026) and the largest with 95 amino acids prior to the motif M1 DLAR (1402–2029). As is evident from the Western blot experiments and the enzymatic assays (Fig. 3a and b; also see Supplementary Tables 2 and 4), the optimal construct is the one with a 45 residue long polypeptide extension that contains the two predicted α-helices. Although the functional significance of this observation is not clear, we note that several splice variants have been mapped on the polypeptide preceding the motif M1 of the vertebrate homologue of DLAR, human hLAR (74% sequence identity to DLAR). Another important observation in this context is that many PTPs viz., PTPH1, PTP1C, SHP1 or SHP2 do not have either of these α helices in the N-terminal region (supplementary Table 5). In the case of other PTP domains, for example PTPβ, this region contains a β-strand in place of the two α helices. That this β-strand plays a conformational role is evident from the crystal structure(s) where it is seen to stack with another β-strand from the PTP catalytic domain (1YGR,

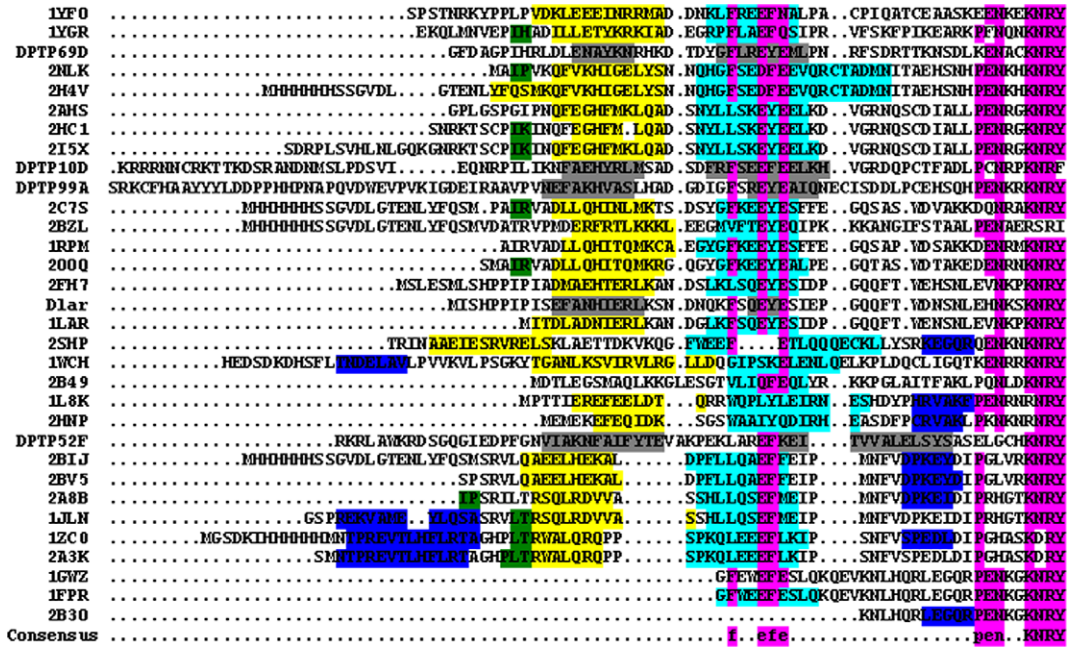

Fig. 5. Sequence and structural diversity in the polypeptide segment preceding the M1 motif of the PTP catalytic domain. Conserved stretches are shown in pink. β strands are highlighted in green. The location of helices α1′ and α2′ are based on the crystal structure of RPTPα in yellow and cyan respectively. While helices other than α1′ and α2′ (seen in PTP structures where the N-terminal extension is present) is in blue, the predicted helices in this region of the *Drosophila* PTPs are highlighted in gray.

2H4V, 2HC1, 2C7s, 1JLN). Thus, given the diversity in sequence and secondary structure of this region, it was not possible to circumvent experimenting with the size of the N-terminal extensions to arrive at a construct that was best suited for over-expression in *E. coli*.

In conclusion, this study provides an experimental data-set of expression constructs that have been examined to provide the most suitable construct for over-expression of a recombinant *Drosophila* PTP in *E. coli*. Although this study examines the case of a specific class of proteins, the results of this analysis could serve as a basis for bioinformatics tools to predict gene constructs that are most likely to yield soluble protein upon expression in *E. coli*.

Acknowledgments

We thank Prof. Kai Zinn for the generous gift of the cosmids that encoded for the *Drosophila* protein tyrosine phosphatases. This work was funded in part by a grant from the Department of Science and Technology, Government of India. L.L.M. is a Junior Research Fellow of the Council for Scientific and Industrial Research, India. B.G. is an International Senior Research Fellow of the Wellcome Trust, United Kingdom.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.pep.2007.10.002.

References

[1] B. Doray, C.D. Chen, B. Kemper, N-terminal deletions and His-tag fusions dramatically affect expression of cytochrome p450 2C2 in bacteria, Arch. Biochem. Biophys. 393 (2001) 143–153.
[2] T. Oswald, W. Wende, A. Pingoud, U. Rinas, Comparison of N-terminal affinity fusion domains: effect on expression level and product heterogeneity of recombinant restriction endonuclease EcoRV, A l Microbiol. Biotechnol. 42 (1994) 73–77.
[3] V. Ramachandiran, C. Willms, G. Kramer, B. Hardesty, Fluorophores at the N terminus of nascent chloramphenicol acetyltransferase peptides affect translation and movement through the ribosome, J. Biol. Chem. 275 (2000) 1781–1786.
[4] E.E. Boeggeman, B. Ramakrishnan, P.K. Qasba, The N-terminal stem region of bovine and human beta 1,4-galactosyltransferase I increases the in vitro folding efficiency of their catalytic domain from inclusion bodies, Protein Expr. Purif. 30 (2003) 219–229.
[5] S.P. Sati, S.K. Singh, N. Kumar, A. Sharma, Extra terminal residues have a profound effect on the folding and solubility of a *Plasmodium falciparum* sexual stage-specific protein over-expressed in *Escherichia coli*, Eur. J. Biochem. 269 (2002) 5259–5263.
[6] M. Hammarstrom, N. Hellgren, S. van Den Berg, H. Berglund, T. Hard, Rapid screening for improved solubility of small human proteins produced as fusion proteins in *Escherichia coli*, Protein Sci. 11 (2002) 313–321.
[7] R.C. Stevens, Design of high-throughput methods of protein production for structural biology, Structure 8 (2000) R177–R185.
[8] S. Idicula-Thomas, P.V. Balaji, Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in *Escherichia coli*, Protein Sci. 14 (2005) 582–592.
[9] M.R. Dyson, S.P. Shadbolt, K.J. Vincent, R.L. Perera, J. McCafferty, Production of soluble mammalian proteins in *Escherichia coli*: identification of protein features that correlate with successful expression, BMC Biotechnol. 4 (2004) 32.
[10] K.G. Johnson, D. Van Vactor, Receptor protein tyrosine phosphatases in nervous system development, Physiol. Rev. 83 (2003) 1–24.

[11] C.A. O'Callaghan, J. Tormo, B.E. Willcox, C.D. Blundell, B.K. Jakobsen, D.I. Stuart, A.J. McMichael, J.I. Bell, E.Y. Jones, Production, crystallization, and preliminary X-ray analysis of the human MHC class Ib molecule HLA-E, Protein Sci. 7 (1998) 1264–1266.
[12] A. Steinle, P. Li, D.L. Morris, V. Groh, L.L. Lanier, R.K. Strong, T. Spies, Interactions of human NKG2D with its ligands MICA, MICB, and homologs of the mouse RAE-1 protein family, Immunogenetics 53 (2001) 279–287.
[13] M.K. McGuffey, K.L. Epting, R.M. Kelly, E.A. Foegeding, Denaturation and aggregation of three alpha-lactalbumin preparations at neutral pH, J. Agric. Food Chem. 53 (2005) 3182–3190.
[14] D.T. Jones, Protein secondary structure prediction based on position-specific scoring matrices, J. Mol. Biol. 292 (1999) 195–202.
[15] R.M. Williams, Z. Obradovi, V. Mathura, W. Braun, E.C. Garner, J. Young, S. Takayama, C.J. Brown, A.K. Dunker, The protein non-folding problem: amino acid determinants of intrinsic order and disorder, Pac Symp. Biocomput. (2001) 89–100.
[16] J.N. Andersen, O.H. Mortensen, G.H. Peters, P.G. Drake, L.F. Iversen, O.H. Olsen, P.G. Jansen, H.S. Andersen, N.K. Tonks, N.P. Moller, Structural and evolutionary relationships among protein tyrosine phosphatase domains, Mol. Cell. Biol. 21 (2001) 7117–7136.
[17] A.M. Bilwes, J. den Hertog, T. Hunter, J.P. Noel, Structural basis for inhibition of receptor protein-tyrosine phosphatase-alpha by dimerization, Nature 382 (1996) 555–559.
[18] A. Salmeen, J.N. Andersen, M.P. Myers, N.K. Tonks, D. Barford, Molecular basis for the dephosphorylation of the activation segment of the insulin receptor by protein tyrosine phosphatase 1B, Mol. Cell. 6 (2000) 1401–1412.
[19] J. King, C. Haase-Pettingell, A.S. Robinson, M. Speed, A. Mitraki, Thermolabile folding intermediates: inclusion body precursors and chaperonin substrates, FASEB J. 10 (1996) 57–66.
[20] B.A. Chrunyk, R. Wetzel, Breakdown in the relationship between thermal and thermodynamic stability in an interleukin-1 beta point mutant modified in a surface loop, Protein Eng. 6 (1993) 733–738.
[21] B.A. Chrunyk, J. Evans, J. Lillquist, P. Young, R. Wetzel, Inclusion body formation and protein stability in sequence variants of interleukin-1 beta, J. Biol. Chem. 268 (1993) 18053–18061.
[22] M. Vedadi, F.H. Niesen, A. Allali-Hassani, O.Y. Fedorov, P.J. Finerty Jr., G.A. Wasney, R. Yeung, C. Arrowsmith, L.J. Ball, H. Berglund, R. Hui, B.D. Marsden, P. Nordlund, M. Sundstrom, J. Weigelt, A.M. Edwards, Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination, Proc. Natl. Acad. Sci. USA 103 (2006) 15835–15840.
[23] N.V. Fedurkina, L.V. Belousova, L.G. Mitskevich, H.M. Zhou, Z. Chang, B.I. Kurganov, Change in kinetic regime of protein aggregation with temperature increase. Thermal aggregation of rabbit muscle creatine kinase, Biochemistry (Moscow) 71 (2006) 325–331.
[24] G.D. Davis, C. Elisee, D.M. Newham, R.G. Harrison, New fusion protein systems designed to give soluble expression in *Escherichia coli*, Biotechnol. Bioeng. 65 (1999) 382–388.
[25] V.N. Uversky, J.R. Gillespie, A.L. Fink, Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41 (2000) 415–427.
[26] A. Ikai, Thermostability and aliphatic index of globular proteins, J. Biochem. (Tokyo) 88 (1980) 1895–1898.
[27] P. Tompa, Intrinsically unstructured proteins, Trends Biochem. Sci. 27 (2002) 527–533.
[28] L. Ni, J. Zhou, T.D. Hurley, H. Weiner, Human liver mitochondrial aldehyde dehydrogenase: three-dimensional structure and the restoration of solubility and activity of chimeric forms, Protein Sci. 8 (1999) 2784–2790.
[29] J. Eswaran, J.E. Debreczeni, E. Longman, A.J. Barr, S. Kna, The crystal structure of human receptor protein tyrosine phosphatase ka a phosphatase domain 1, Protein Sci. 15 (2006) 1500–1505.
[30] J.S. Zhang, F.M. Longo, LAR tyrosine phosphatase receptor: alternative splicing is preferential to the nervous system, coordinated with cell growth and generates novel isoforms containing extensive CAG repeats, J. Cell Biol. 128 (1995) 415–431.