Soluble expression of proteins correlates with a lack of positively-charged surface

Pedro Chan¹*, Robin A. Curtis² & Jim Warwicker¹

¹Faculty of Life Sciences, ²School of Chemical Engineering and Analytical Sciences, University of Manchester, Manchester Institute of Biotechnology, 131 Princess Street, Manchester M1 7DN, UK.

Prediction of protein solubility is gaining importance with the growing use of protein molecules as therapeutics, and ongoing requirements for high level expression. We have investigated protein surface features that correlate with insolvability. Non-polar surface patches associate to some degree with insolvability, but this is far exceeded by the association with positively-charged patches. Negatively-charged patches do not separate insoluble/soluble subsets. The separation of soluble and insoluble subsets by positive charge clustering (area under the curve for a ROC plot is 0.85) has a striking parallel with the separation that delineates nucleic acid-binding proteins, although most of the insoluble dataset are not known to bind nucleic acid. Additionally, these basic patches are enriched for arginine, relative to lysine. The results are discussed in the context of expression systems and downstream processing, contributing to a view of protein solubility in which the molecular interactions of charged groups are far from equivalent.

Protein solubility and propensity to aggregate has been central to biotechnology and biosciences through the era of recombinant protein expression. It is also becoming increasingly important in the area of formulation and preparation of biologics (therapeutic proteins), and in consideration of disorders arising from misfolding¹. A common view of protein aggregation at relatively high concentration holds that partial unfolding (a structural feature) leads to association of non-polar stretches of amino acids (a sequence feature). Whilst structural and sequence properties combine in this view of protein aggregation, computational algorithms that attempt to predict solubility largely divide into those based on sequence and those based on structure, although some features (e.g. net charge) span this division.

A key question relates to how solubility is defined in benchmark sets. Early work² distinguishes between proteins that form inclusion bodies (IBs) and those that do not, with a study of sequence features. The two properties correlating best with IB formation were found to be average charge (more net charge, less IB), and turn-forming residue fraction (more gives IBs, perhaps due to slow folding e.g. with prolines). Other work³ also uses the IB/non-IB distinction, together with sequence and structure-based correlations with solubility, including thermostability and relative lack of β-sheet. Some reports define soluble proteins as those for which a structure has been solved and deposited in the protein data bank, PDB⁴. This definition is used alongside resources that record progress in protein expression for structural genomics⁵,⁶, such as the TargetDB database⁷. Machine-learning techniques are then employed to optimise distinction between soluble and insoluble proteins, although it can be difficult to extract physico-chemical interpretation from such methods. Other work combines machine-learning with soluble/insoluble datasets obtained through keyword searching in the literature⁸. The relationship between mRNA levels and protein solubility in E. coli has been examined⁹. Proteins with sequence more prone to aggregation are generally expressed at lower levels, where amino acid polarity is used to indicate aggregation potential, i.e. proteins with a more non-polar sequence have lower mRNA levels. The REFOLD database¹⁰ annotates proteins as soluble or insoluble, but in practice all of these proteins have been expressed through IB formation.

There are a number of aggregation prediction schemes based on the experimental observation that many proteins can be induced to adopt an amyloid, β-rich conformation¹¹. These include TANGO¹², PASTA¹³, and Zyggregator¹⁴. Such schemes can include many factors, but generally, the β-forming propensity for linear segments of amino acid sequence is an important element. A 3D surface polarity approach has been adopted in the redesigning of protein surface to improve solubility¹⁵, with the introduction of groups to break up non-polar patches. This is reminiscent of the discovery that charges on the surfaces of hyperthermophile proteins are more closely packed, on average, than those in mesophile proteins¹⁶. It was assumed that the higher temperature environment of hyperthermophiles increases the strength of hydrophobic interactions, leading to the requirement for a more stringent breaking up of non-polar patches with charges. A charge influence has also appeared in
the context of translation rate, a property that will impact on protein production and therefore potentially solubility. A dependence of ribosomal velocity on positively charged residues in newly synthesised proteins has been found, due to interaction with the negatively-charged ribosomal exit tunnel 17. More generally, translation rate has a well-studied correlation with codon bias 18.

Methods for predicting protein solubility have been reviewed 19. The availability of experimental data, where proteins have been expressed in consistent conditions, continues to present a significant problem with assessing prediction schemes. A significant study addressing this point used a high throughput cell-free system for classification of E. coli protein solubility 20. The authors of this work concluded that factors correlating to some degree with solubility include charge and structural class, whilst algorithms based largely on propensity to form β-structure/amyloid performed less well, although a machine-learning study subsequently identified a correlation between sequence-based calculation of physico-chemical properties and measured solubility for this dataset 21.

In the current work, computational methods for characterising charge and potential distributions in proteins 22 have been used alongside patch-based calculations of surface properties 23 to analyse the properties of soluble and insoluble subsets of proteins. The experimental data used in this study derive from cell-free expression using the PURE system of E. coli factors, lacking chaperones 24. Encouraged by a study in which computation over many proteins revealed a correlation between electrostatic properties and subcellular location 25, a similar approach was used in respect of solubility. Whilst some correlation is found between insolubility and larger non-polar patches, by far the most significant relationship associates insolubility with large positively-charged patches. The pattern underlying this unexpected result is similar to that which separates nucleic acid (NA)-binding from non-NA-binding proteins.

**Results**

**Surface potential patches and solubility.** At neutral pH most of the insoluble and soluble dataset proteins are predicted to be moderately negatively-charged, and there is no significant separation of the distributions (Fig. 1a, p = 0.872 for a Mann-Whitney test of subsets being sampled from the same underlying distribution). The maximal positive and negative potential patches for each protein show quite different behaviour, with no significant separation for negative potential (p = 0.227), but clear separation for positive potential (p = 7.1 × 10⁻¹³, Fig. 1b). A patch analysis of charge clustering (with 13 Å patch radius) was performed in order to establish whether the positive potential patches, based on contours, were mirrored in charge geometry. This is the case, with the largest net positive charge on a patch also distinguishing soluble and insoluble protein datasets (p = 2.2 × 10⁻⁴, Fig. 1c).

**Surface polarity and solubility.** We next examined a potential role for the association of proteins via non-polar surfaces, through calculation of non-polar to polar solvent accessible surface area (SASA) ratios, for patches of radius 13 Å centred on each atom. The maximum of this ratio was identified for each protein. Fig. 1d shows the separation of soluble and insoluble subsets (p = 2.3 × 10⁻³). Whilst there is some correlation between increased non-polarity and insolubility, it is far smaller than that exhibited by positive potential. ROC plot analysis demonstrates this distinction (Fig. 2). An area under the curve (AUC) of 0.85 for the positive potential features (Fig. 2a) compares with an AUC of 0.62 for the

![Figure 1](https://www.nature.com/scientificreports/srep03333/figure1.png)
protein aggregation in the cell free expression system 20. If this is also positively-charged or negatively-charged at neutral pH separates positive and negative charge, arginine and lysine 0.138, not shown).

Positive and negative charge, arginine and lysine. Sequence-based calculation of the fraction of charged groups that are either positively-charged or negatively-charged at neutral pH separates soluble and insoluble subsets (p = 1.417 × 10⁻⁹, not shown). Consistent with the patch calculations, a higher fraction of positive charge tends towards insolubility. With a greater separation of soluble and insoluble subsets for the (3D) patch-based property, relative to the sequence-based charge fraction, the structural property appears to be a crucial component in a physico-chemical understanding of the cell-free expression data.

Discussion
Our results indicate that factors contributing on average to separation of the structurally annotated soluble and insoluble subsets in cell-free expression 22, are non-polar surface (moderate contribution), and positively-charged patches (major contribution, particularly where Arg is more prevalent than Lys). Correlation between largest positive patch and insolubility implies that this property, or another feature to which it is strongly related, acts in some direct or indirect way to promote protein-protein interactions. It could be argued that a concentration of positive charge may tend towards lower folded state stability through unfavourable charge interactions, and thus influence solubility via (partial) unfolding. However, a similar influence would be expected for negatively-charged patches, which is absent. Through what other mechanisms could positive charge clustering contribute to insolubility? Given that the characteristic for insolubility observed in the current work closely corresponds to 3 close salt-bridges or greater than 3 flexible +/− charge interactions, consistent with a net charge threshold of about 4.5 (Fig. 1c). Briefly, having estimated maximal concentrations of charged protein (positive) and NA (negative) sites each at 4 mM, concentration ramps up to the these values are used to account for subsets of protein and NA sites possessing the appropriate...
coli E. small fraction of inclusion body material for proteins expressed in

case of direct protein-protein interactions.

concentrations of interacting components, and thus also applies to
temporal complexes. The diagonal drawn on Fig. 3a relates to equal

relatively high concentrations, leads to a substantial fraction of (trans-

cy, independent of the nature of the precipitant 44. These experi-

ments, which reflect protein-protein interactions between folded

(patch calculations for these 7 proteins. No cor-

unshielded charge (see Methods for more detail). This simple cal-

culation (Fig. 3a) shows that a weak interaction, coupled with

relatively high concentrations, leads to a substantial fraction of (tran-

cient) complexes. The diagonal drawn on Fig. 3a relates to equal

concentrations of interacting protein sites (vertically). The heat map shows the proportion of

interacting protein sites that are complexed (scale bar under the map). See

text for more detail. (b) A hypothetical scheme is drawn in which protein-

NA interactions are mediated by charge interactions (upper left), followed

by partial unfolding concomitant with NA base – protein interactions

be moderated unless functionally associated with nucleic acid

binding, perhaps to avoid pathways such as that hypothesised in

Fig. 3b. Generally, NA-binding proteins such as transcription factors

can be difficult to express 46, 47.

It is worth stating the fundamental points that relate protein sur-

face charges to protein solubility. The hydration of charged groups is

related more to the avoidance of basic clusters (and perhaps NA

binding), whereas the purified protein experiments may be probing,

in part, a more general stickiness associated with the Arg sidechain.

The authors of the precipitant-based solubility study concluded that

strong water binding by acidic amino acids may underpin the

results 44. The current work

suggests that surface charge chaperoning may be a contributing factor. A growing body of data is becoming available with which to test

such hypotheses 46.

Solubility measurements for 7 proteins in different precipitants

show that negative surface charge correlates with increased solubil-

ity, independent of the nature of the precipitant 44. These experi-

ments, which reflect protein-protein interactions between folded

(purified) proteins, are quite different to the cell-free translation

study 20 on which the current work is based, but given the importance

of charge, we made patch calculations for these 7 proteins. No cor-

relation is seen for maximum positive patch size and solubility (R =

0.393, p = 0.191, not shown), but a relationship may be present

between overall Lys to Arg ratio and solubility (R = 0.720, p =

0.034, not shown), with (again) a higher ratio tending towards more

soluble. One interpretation is that the Arg sidechain is particularly

prone to interactions. Solubility in the cell-free system could be

related more to the avoidance of basic clusters (and perhaps NA

binding), whereas the purified protein experiments may be probing,

in part, a more general stickiness associated with the Arg sidechain.

The authors of the precipitant-based solubility study concluded that

strong water binding by acidic amino acids may underpin the

results 44. The balance between negative and positive amino acid side-

chain charges in fine-tuning solubility, remains to be established.

It is of interest that Arg has been identified as a ubiquitous interacting

amino acid in informatics studies, with an elevated propensity (relative

to average surface occurrence) for interfaces in both protein-protein and

Figure 3 | Weak interactions and association in a crowded environment. (a) Two species interact with an energy of 15 kJ/mole. Concentrations are varied (0 to 4 mM) for protein interacting sites (horizontally) and NA interacting sites (vertically). The heat map shows the proportion of interacting protein sites that are complexed (scale bar under the map). See text for more detail. (b) A hypothetical scheme is drawn in which protein-NA interactions are mediated by charge interactions (upper left), followed by partial unfolding concomitant with NA base – protein interactions (upper right), then protein-protein association through non-polar interactions (lower right), and finally dissociation of protein from NA (lower left).
protein-NA complexes. Cation – π interactions, involving Arg, are common at protein-protein interfaces, and Arg is also common in protein crystal contacts at low ionic strength. Arginine content of antigen-combining sites in antibodies is correlated with increased non-specific binding. The excipient properties of Arg are also of interest. Solutions of Arg can be effective in solubilising proteins, an effect that becomes more pronounced in mixtures with Glu. This solubility enhancement is related to an increase in the number of Arg and Glu molecules forming interactions with the protein. The interacting properties of Arg cover a range of systems. We suggest that such diversity may lead to a correlation of Arg enrichment with insolvibility, whether clustering into patches (with Lys) for polyanion binding, or more generally over a protein surface.

Reduction of positive patches should be of use as a design tool for expression systems, and disrupting Arg with other charges could aid the maintenance of high concentrations of purified protein in solution. The hypothesis of protein basic patch interactions with NA in expression systems could be investigated with uncoupling of transcription and translation, to vary relative mRNA and protein levels. There is much yet to establish about the association between basic clusters, and Arg enrichment, and insolubility, given for example the report that green fluorescent protein engineered to bear high net positive or negative charge expresses in E. coli and is much more soluble than wild-type protein. In this case perhaps the extreme net charges provide sufficient repulsive interactions to overcome other effects.

**Methods**

**Soluble and insoluble datasets and DNA-binding/non-binding protein datasets.** Subsets for soluble and insoluble *E. coli* protein expression in the cell-free system were defined following the authors’ description. Specifically, soluble proteins are those with a solubility of more than 70%, and insoluble with a solubility of less than 30%. Percentage solubilities had been obtained, following cell-free expression of radiolabelled protein, as the ratio of soluble protein (supernatant from a solution. The hypothesis of protein basic patch interactions with NA in expression systems could be investigated with uncoupling of transcription and translation, to vary relative mRNA and protein levels. There is much yet to establish about the association between basic clusters, and Arg enrichment, and insolubility, given for example the report that green fluorescent protein engineered to bear high net positive or negative charge expresses in E. coli and is much more soluble than wild-type protein. In this case perhaps the extreme net charges provide sufficient repulsive interactions to overcome other effects.

Soluble and insoluble protein datasets were available for processing.

Sets of DNA-binding and non-binding proteins were obtained from earlier work. Many of these PDB ids were accessible and ran successfully through the electrostatic potential patch analysis (128 DNA-binding proteins, 108 non-binding). Calculations were also made for a set of 7 proteins for which solubility data were available in precipitant studies, using the same PDB ids specified in that work.

**Charge, potential and polarity calculations.** For polarity analysis, a sphere of radius 13 Å was centred on each non-hydrogen atom. Polar and non-polar solvent accessible surface area was then summed for all non-hydrogen atoms within that sphere, using a 1.4 Å radius solvent probe and polar/non-polar character assigned according to atom type and functional group. The relative polarity of a patch is then calculated as the ratio of non-polar SASA to polar SASA, and the maximum value of this ratio (i.e. most non-polar region) recorded for each protein. When the polar and non-polar SASAs are summed for each patch, the average of this distribution over patches is about 3000 Å^2. In comparison, a typical interface between proteins buries about 1600 Å^2 in total, although this is quite variable. Considering that the entirety of each of the two contributing surfaces will not be buried in an association, a patch radius of 13 Å seems reasonable in generating a footprint for non-specific protein-protein interactions.

Electrostatic potential was calculated around each protein using a Finite Difference Poisson-Boltzmann methodology, with negatively-charged Asp, Glu sidechains and C-termini, and positively-charged Lys, Arg and N-termini. Ionic strength was 0.15 M in the calculations. The resulting potential map was contoured at thresholds of ±1.0 kT/e. Importantly, the contours were drawn on a single shell of the calculation grid (on the solvent side of the interface), so that the number grid points in each contoured patch effectively represent the size of that patch. Grid size for electrostatic potential calculation was a constant 0.6 Å, independent of protein. A parallel approach was introduced to confirm that positive charge location underpins the contours of positive potential. The patch analysis described for surface polarity was also used to record the maximum net charge within a geometrical patch. For this purpose, sidechain charges were approximated at Cβ atoms to minimise the effects of sidechain conformational variation.

Receiver Operator Characteristic (ROC) plots were generated for the ability of calculated features to discriminate between soluble and insoluble proteins. As the numerical value of a feature is varied and applied as a threshold to the datasets, corresponding true positive rates (TPRs) and false positive rates (FPRs) are calculated and given in a ROC plot. Area Under the Curve (AUC) is used to estimate effectiveness for separating datasets, with 1.0 equating to complete separation and 0.5 to random. The Mann-Whitney U test was applied to the calculated feature subsets. The probability of occurrence of these particular feature values, if there is no difference in the underlying distributions, is given. A significant difference is inferred if this probability is < 0.05.

To investigate whether a relationship exists between calculated features and expression at the mRNA level, protein IDs for soluble and insoluble datasets were mapped to mRNA abundances for *E. coli* proteins.

A model for non-specific protein-nucleic acid charge interactions. The model of Fig. 3a is based on a 15 kJ/mole interaction, or about 3 salt-bridges, since typical pKa shifts for a surface salt-bridge are 1 pK unit or 5–6 kJ/mole. Total concentration of protein was calculated for an estimate of 2.35 × 10^6 protein molecules in an *E. coli* cell of 1 μm, giving 4 mM. Summing the estimated contributions of RNAs and mRNAs and comparing with protein molecular weight, gives a ratio of about 1:10, nucleic acid to protein. Each protein molecule, of average molecular weight 60 kDa, might bind to a polyanion through a single positive patch, whereas each polyanion nucleic acid molecule has multiple binding sites. With a single base molecular weight of about 0.32 kD, assuming that binding sites could recur approximately every 12 bases, then two factors of 10 approximately cancel and the maximum concentrations of oppositely-charged sites are roughly equal. Although 4 mM is set as this maximum, only a subset of proteins will exhibit positive patches above a certain threshold, whilst there are many factors that will contribute to structuring of nucleic acids and the masking of negative charge. Thus linear concentration ramps, up to 4 mM, are applied in Fig. 3a for interacting subsets of protein and nucleic acid. The map is then generated as the proportion of the interacting subset of proteins that is bound to nucleic acid, as the concentrations are altered, given the 15 kJ/mole interaction.
21. Agostini, F., Vendruscolo, M. & Tartaglia, G. G. Sequence-based prediction of protein solubility. *J Mol Biol* **421**, 237–241 (2012).

22. Warwick, J. Continuum dielectric modelling of the protein-solvent system, and calculation of the long-range electrostatic field of the enzyme phosphoglycerate mutase. *J Theor Biol* **121**, 199–210 (1986).

23. Cole, C. & Warwick, J. Side-chain conformational entropy at protein-protein interfaces. *Proteins* **11**, 2860–2870 (2002).

24. Shimizu, Y. et al. Cell-free translation reconstituted with purified components. *Nat Biotechnol* **19**, 751–755 (2001).

25. Chan, P. & Warwick, J. Evidence for the adaptation of protein pH-dependence to subcellular pH. *BMC Biol* **7**, 69 (2009).

26. Warwick, J., Engelmann, B. P. & Steitz, T. A. Electrostatic calculations and model-building suggest that DNA bound to CAP is sharply bent. *Proteins* **2**, 283–289 (1987).

27. Chen, Y. C., Wright, J. D. & Lim, C. DR_bind: a web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry. *Nucleic Acids Res* **40**, W249–256 (2012).

28. Swinger, K. K. & Rice, P. A. IHF and HU: flexible architects of bent DNA. *Biochim Biophys Acta* **1784**, 600–610 (2008).

29. de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M. & Vogel, C. Global signatures of protein and mRNA expression levels. *Mol Biosyst* **5**, 1512–1526 (2009).

30. Taniguchi, Y. et al. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–538 (2010).

31. Bernstein, J. A., Khodursky, A. B., Lin, P. H., Lin-Chao, S. & Cohen, S. N. Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc Natl Acad Sci USA* **99**, 9697–9702 (2002).

32. Valax, P. & Georgiou, G. Molecular characterization of beta-lactamase inclusion bodies produced in Escherichia coli. 1. Composition. *Biotechnol Prog* **9**, 539–547 (1993).

33. Maachupalli-Reddy, J., Kelley, B. D. & De Bernardez Clark, E. Effect of inclusion body contaminants on the oxidative renaturation of hen egg white lysozyme. *Biotechnol Prog* **13**, 144–150 (1997).

34. Sedlak, E., Fedunova, D., Vesela, V., Sedlakova, D. & Antalik, M. Polyionan hydrophobicity and protein basicity affect protein stability in protein-polyanion complexes. *Biomacromolecules* **10**, 2533–2538 (2009).

35. Cawley, A. & Warwick, J. eIIAe-binding protein regulation of mRNAs with differential 5'–UTR secondary structure: a polyelectrolyte model for a component of protein-mRNA interactions. *Nucleic Acids Res* **40**, 7666–7675 (2012).

36. Mosakowska, D. E. Expression of nuclear hormone receptors in Escherichia coli. *Curr Opin Biotechnol* **9**, 502–505 (1998).

37. Yang, W. C., Welsh, J. P., Lee, J., Cooke, J. P. & Swartz, J. R. Solubility partner IF2 Domain I enables high yield synthesis of transducible transcription factors in Escherichia coli. *Protein Expr Purif* **80**, 145–151 (2011).

38. Durnet, A. C., Chocka, A. M., Kaler, E. W. & Lenhoff, A. M. Effects of pH on protein–protein interactions and implications for protein phase behavior. *Biochim Biophys Acta* **1784**, 600–610 (2008).

39. Neil, B. L., Asthagiri, D., Velez, O. D., Lenhoff, A. M. & Kaler, E. W. Why is the osmotic second virial coefficient related to protein crystallization? *Journal of Crystal Growth* **196**, 377–387 (1999).

40. Dudgeon, K. et al. General strategy for the generation of human antibody variable component of protein-mRNA interactions. *Biomacromolecules* **10**, 2533–2538 (2009).

41. Kvam, E., Siersk, M. R., Shoemaker, C. B. & Messer, A. Physico-chemical determinants of soluble intrabody expression in mammalian cell cytoplasm. *Protein Expr Des* **23**, 489–498 (2010).

42. Wayne, N. & Bolon, D. N. Charge-rich regions modulate the anti-aggregation activity of Hsp90. *J Mol Biol* **401**, 931–939 (2010).

43. Ahn, J. H., Keum, J. W. & Kim, D. M. Expression screening of fusion partners from an E. coli genome for soluble expression of recombinant in a cell-free protein synthesis system. *PLoS One* **6**, e22875 (2011).

44. Kramer, R. M., Shende, V. R., Motl, N., Pace, C. N. & Scholtz, J. M. Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. *Biophys J* **102**, 1907–1915 (2012).

45. Jones, S., van Heyningen, P., Berman, H. M. & Thornton, J. M. Protein-DNA interactions: A structural analysis. *J Mol Biol* **287**, 877–896 (1999).

46. Jones, S., Daley, D. T., Luscombe, N. M., Berman, H. M. & Thornton, J. M. Protein-RNA interactions: a structural analysis. *Nucleic Acids Res* **29**, 943–954 (2001).

47. Crowley, P. B. & Golovin, A. Cation-pi interactions in protein-protein interfaces. *Proteins* **59**, 231–239 (2005).

48. Wayne, N. & Bolon, D. N. Charge-rich regions modulate the anti-aggregation activity of Hsp90. *J Mol Biol* **287**, 877–896 (1999).

49. Bir tallan, S. et al. The intrinsic contributions of tyrosine, serine, glycine and arginine to the affinity and specificity of antibodies. *J Mol Biol* **377**, 1518–1528 (2008).

50. Golovanov, A. P., Hautbergue, G. M., Wilson, S. A. & Lian, L. Y. A simple method for improving protein solubility and long-term stability. *J Am Chem Soc* **126**, 8933–8939 (2004).

51. Koo, T. Y. & Park, T. H. Expression of recombinant human growth hormone in a soluble form in Escherichia coli by slowing down the protein synthesis rate. *J Microbiol Biotechnol* **17**, 579–585 (2007).

52. Lawrence, M. S., Phillips, K. J. & Liu, D. R. Supercharging proteins can impart unusual resilience. *J Am Chem Soc* **129**, 10110–10112 (2007).

53. Bairoch, A. et al. The Universal Protein Resource (UniProt). *Nucleic Acids Res* **33**, D154–159 (2005).

54. Wang, G. & Dunbrack Jr, R. L. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res* **33**, W94–99 (2005).

55. Ahmad, S. & Sarai, A. Moment-based prediction of DNA-binding proteins. *J Mol Biol* **314**, 65–71 (2004).

56. Radko, G., Shipman, P., Sizov, A., Leslie, C. & Ben-Tal, N. iDBPs: a web server for the identification of DNA binding proteins. *Bioinformatics* **26**, 692–693 (2010).

57. Ahmad, S. & Sarai, A. Moment-based prediction of DNA-binding proteins. *J Mol Biol* **314**, 65–71 (2004).

58. Lo Conte, L., Chothia, C. & Janin, J. The atomic structure of protein-protein recognition sites. *J Mol Biol* **285**, 2177–2178 (1999).

59. Warwick, J., Simplesse method for pKa and acid pH-dependent stability estimation in proteins: removing dielectric and counterion boundaries. *Protein Sci* **8**, 418–425 (1999).

60. Neidhardt, F. C. & Umbarger, H. E. in Escherichia Coli and Salmonella: Cellular and Molecular Biology, 2nd edition, Vol. I, (ed. Neidhardt, F. C.) (American Society of Microbiology, 1996).

**Acknowledgments**

We thank E.A. McKenzie and M.A. Carballo for discussion, and the UK BBSRC for contributing PhD funds (PC).

**Author contributions**

J.W., P.C. and R.A.C. designed the study. P.C. and J.W. performed the computational analyses. J.W. and R.A.C. wrote the manuscript, which was reviewed by all authors.

**Additional information**

Competing financial interests: The authors declare no competing financial interests.

*How to cite this article:* Chan, P., Curtis, R.A. & Warwick, J. Soluble expression of proteins correlates with a lack of positively-charged surface. *Sci. Rep.* **3**, 3333; DOI:10.1038/srep03333 (2013).